生物医学类基础参考书

生物信息学
理论与医学实践

BIOINFORMATICS THEORY AND MEDICAL PRACTICE

主 编 李 霞　　副主编 张 岩 陈秀杰 陈丽娜



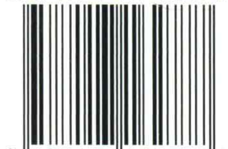
人民卫生出版社

策划编辑　戴薇薇
责任编辑　李海凌　戴薇薇
封面设计　永诚天地　代珊珊
版式设计　邹桂荣

生物信息学理论与医学实践

BIOINFORMATICS THEORY AND MEDICAL PRACTICE

销售分类 / 生物信息


ISBN 978-7-117-16277-7
定　价：158.00 元

人卫社官网　www.pmph.com　出版物查询，在线购书
人卫医学网　www.ipmph.com　医学考试辅导，医学数据库服务，医学教育资源，大众健康资讯

BIOINFORMATICS THEORY AND MEDICAL PRACTICE

生物信息学理论与医学实践

主　　编　李　霞

副主编　张　岩　陈秀杰　陈丽娜

编　　委（以姓氏笔画排序）

王艳秋　刘　磊　李　晋　李　霞

李春权　肖　云　汪强虎　张　岩

张绍军　陈　曦　陈丽娜　陈秀杰

陈晓文　周　猛　姜　伟　宫滨生

徐良德　徐建凯　智　慧

学术秘书　王　宏

人民卫生出版社

图书在版编目（CIP）数据

生物信息学理论与医学实践/李霞主编.—北京：人民卫生出版社，2013.1

ISBN 978-7-117-16277-7

Ⅰ.①生… Ⅱ.①李… Ⅲ.①生物信息论 Ⅳ.①Q811.4

中国版本图书馆CIP数据核字（2012）第283419号

人卫社官网 www.pmph.com 出版物查询，在线购书
人卫医学网 www.ipmph.com 医学考试辅导，医学数据库服务，医学教育资源，大众健康资讯

版权所有，侵权必究！

生物信息学理论与医学实践

主　　编：李　霞
出版发行：人民卫生出版社（中继线 010-59780011）
地　　址：北京市朝阳区潘家园南里19号
邮　　编：100021
E - mail：pmph @ pmph.com
购书热线：010-67605754 010-65264830
010-59787586 010-59787592
印　　刷：三河市宏达印刷有限公司
经　　销：新华书店
开　　本：787×1092 1/16 印张：35
字　　数：852千字
版　　次：2013年1月第1版 2013年1月第1版第1次印刷
标准书号：ISBN 978-7-117-16277-7/R·16278
定　　价：158.00元
打击盗版举报电话：010-59787491 E-mail：WQ @ pmph.com
（凡属印装质量问题请与本社销售中心联系退换）

序 言

20世纪90年代初，人类基因组计划（human genome project，HGP）的启动推动了生物学、医学、药学与信息科学之间的紧密联系，海量数据的收集、存储、分析及解释促使全世界科学家思考生物学、医学和药学发展的新思路，生物信息学就是在这样的背景下产生并蓬勃发展起来的。生物信息学（bioinformatics）是以数理科学为理论基础、以计算机技术为工具，进行深层次生物医学海量数据挖掘与分析的多学科交叉的新兴学科。伴随着现代生物技术的发展，生物信息学在现代生物学、医学和药学的发展中发挥着重要作用。

随着新一代测序技术、生物芯片技术、药物筛选技术等的快速发展，现代生物学、医学、药学研究已经由单一因素、单个分子层面进入到高通量、大规模的组学研究时代。面对信息含量大、数理逻辑强的生物学、医学、药学资源，传统的实验方法遇到巨大挑战，已经难以独立解决众多复杂的生物学、医学、药学问题。在此基础上，以海量数据分析为研究内容的生物信息学逐渐成为生物学、医学、药学研究领域不可或缺的组成部分。生物信息学理论能够广泛地应用于生物学、医学、药学等研究领域，如人类疾病病因学研究、临床诊断标志物识别、疾病分型和预后预测、遗传调控机制和分子通路建立、药物靶标识别与药物设计、新兴生物大分子发现与功能分析、生理模拟与病理推断、动植物育种与分子进化等方面，并能够极大地促进个性化医疗的发展。由此，我们总结多年积累的学术思想、研究心得及结果，编著了《生物信息学理论与医学实践》一书。本书旨在为生物学、医学、药学研究领域的科学工作者及生物信息学领域的同行、学生等人员介绍生物信息学基础理论、数据分析方法及其在生物学、医学、药学领域中的应用成果。

本书不仅对生物信息学研究领域的基础知识及基本理论进行了详细介绍，

如网络数据资源、序列比对、分子进化、基因芯片及蛋白质结构等；而且介绍了目前国内外生命科学研究应用的热门技术及热点领域，如新一代测序技术、富集分析技术、表观遗传学分析技术、microRNA与疾病研究及药物靶点筛查技术，并对书中涉及的各种分析技术给出详细的应用实例。我们希望能与感兴趣的读者交流，有机会完善本书。限于作者目前的水平，加之时间仓促，书中必有许多不足之处，希望能够得到读者的指正。

本书每一章的编者都有相关领域多年并丰富的研究经历，每一章都凝聚了他们的学术思想及科研成果。他们在百忙之中精心组织材料、字斟句酌编写本书，在此我们对全体编者的无私奉献表示衷心的感谢！多年来，我们的工作得到了哈尔滨医科大学各方面的大力支持与热情鼓励，同时也得到了国家自然科学基金的支持，谨在此一并表示诚挚的谢意！

李　霞

2012年12月

目 录

绪 论

INTRODUCTION TO BIOINFORMATICS MOLECULAR

第一节 生物信息学的产生及发展 / 02

一、生物信息学的产生 / 02

二、生物信息学的发展历史 / 03

三、生物信息学在未来生命科学研究中的作用 / 06

第二节 生物信息学的主要研究内容 / 07

一、基因组信息学 / 07

二、转录组信息学 / 09

三、蛋白质组信息学 / 11

四、代谢组信息学 / 12

第三节 当前生物信息学研究的热点 / 14

一、新一代测序数据的生物信息学分析 / 14

二、非编码区序列分析与功能识别 / 15

三、整合信息组学 / 15

四、转化医学和临床生物信息学 / 16
五、生物信息学与新药研究 / 16

第一章 序列比对与序列特征分析

CHAPTER 1 SEQUENCE ALIGNMENT AND ANALYSIS OF SEQUENCE CHARACTERISTICS

第一节 获取DNA、RNA和蛋白质序列 / 20
一、DNA序列的获取 / 20
二、RNA序列的获取 / 28
三、蛋白质序列的获取 / 29
第二节 双序列比对 / 32
一、序列比对的相关概念 / 32
二、双序列比对算法 / 37
第三节 多序列比对 / 43
一、多序列比对简介 / 43
二、多序列比对的方法 / 44
三、多序列比对常用工具和数据库 / 48
第四节 核酸序列特征分析 / 56
一、基因开放读码框的识别 / 56
二、内含子/外显子剪切位点的识别 / 58
三、序列模体的查找和可视化工具 / 58
四、密码子使用模式的分析 / 59
五、限制性核酸内切酶位点分析 / 60

六、重复序列的查找 / 61

第五节 蛋白质序列特征分析 / 62

一、蛋白质的理化性质分析 / 62
二、蛋白质的跨膜结构分析 / 64
三、蛋白质信号肽的预测和识别 / 67
四、蛋白质的卷曲螺旋预测 / 67
五、糖基化位点的预测与识别 / 68
六、磷酸化位点的预测与识别 / 68

第六节 应用实例 / 69

一、利用Spidey工具识别mRNA/cDNA的外显子组成 / 69
二、利用Spidey工具进行可变性剪切的分析 / 71

第二章 基因芯片数据分析
CHAPTER 2 MICROARRAY DATA ANALYSIS

第一节 基因芯片平台简介 / 76

一、cDNA芯片 / 76
二、寡核苷酸芯片 / 77
三、原位合成芯片 / 77
四、光纤微珠芯片 / 79

第二节 基因芯片数据的预处理 / 81

一、基因芯片数据的提取 / 81
二、数据过滤 / 81

三、补缺失值 / 82
四、数据对数化处理 / 82
五、数据标准化 / 83
六、应用举例 / 85

第三节 特征基因挖掘 / 89

一、特征选择的过程 / 89
二、特征选择方法的分类 / 92
三、应用举例 / 95

第四节 基因芯片数据的聚类分析 / 98

一、聚类分析中的距离(相似性)尺度函数 / 98
二、常用的聚类方法 / 100
三、应用实例 / 104

第五节 基因芯片数据的分类分析 / 108

一、*k*近邻分类法 / 108
二、决策树 / 108
三、支持向量机 / 110
四、分类器的分类效能评价 / 111

第六节 基因芯片数据库及常用分析软件 / 113

一、基因表达数据库(gene expression omnibus, GEO)/ 113
二、基因芯片显著性分析
(significance analysis of microarray, SAM)/ 114
三、Cluster and TreeView / 114
四、BRB-ArrayTools / 115
五、R语言和Bioconductor / 115
六、Matlab: Bioinformatics Toolbox / 115

第三章　基因注释与功能分类

CHAPTER 3 GENE ANNOTATION AND FUNCTIONAL CLASSIFICATION

第一节　引言 / 118

第二节　基因注释数据库　/　119

一、GO(gene ontology)数据库 / 119
二、KEGG数据库 / 125
三、其他常见生物学通路数据库 / 130

第三节　基因功能预测　/　132

一、基因功能预测的目的和意义 / 132
二、基因功能预测的基本原理 / 133
三、基因功能预测的常用工具 / 136

第四节　基因集合富集分析　/　139

一、富集分析的目的和意义 / 139
二、富集分析的基本原理 / 139
三、富集分析常用工具 / 140
四、富集分析应用实例 / 142

第五节　基因功能比较　/　145

一、基因功能比较的意义 / 145
二、语义相似性原理与算法 / 146
三、生物学术语相似性 / 146
四、基因(基因产物)功能比较 / 148
五、基因集合功能比较 / 149

六、常用工具 / 149

第四章　蛋白质结构分析

CHAPTER 4 PROTEIN STRUCTURE ANALYSIS

第一节　蛋白质高级结构 / 154

一、蛋白质的高级结构特征 / 154

二、蛋白质结构域与家族分类 / 157

三、蛋白质结构可视化软件 / 158

第二节　蛋白质的结构解析与预测 / 162

一、蛋白质结构实验检测技术与结构解析 / 162

二、蛋白质结构数据库 / 164

第三节　蛋白质高级结构预测方法 / 175

一、蛋白质二级结构预测方法及软件 / 175

二、蛋白质三级结构的预测方法及软件 / 180

三、对结构预测结果的评价 / 186

第四节　基于蛋白质结构的功能分析 / 187

一、蛋白质结构与功能基础 / 187

二、蛋白质结构与功能关系数据库 / 188

三、从蛋白质结构推断其功能的方法与分析软件 / 193

四、蛋白质相互作用与蛋白质功能 / 197

第五节　蛋白质的结构异常与疾病 / 202

一、蛋白质序列变化引发疾病 / 202

二、蛋白质折叠错误引发疾病 / 202
三、疾病过程中蛋白质的相互作用 / 203

第六节 应用实例 / 205

一、蛋白质结构信息在类风湿性关节炎致病基因挖掘中的应用 / 205
二、蛋白质结构转换几率与疾病的发生 / 208

第五章 分子进化分析

CHAPTER 5 MOLECULAR EVOLUTION ANALYSIS

第一节 系统发生分析与重建 / 214

一、核苷酸置换模型与氨基酸置换模型 / 214
二、系统发育树重建方法 / 219
三、分子钟假说 / 224

第二节 核苷酸和蛋白质的适应性进化 / 226

一、中性与近中性理论 / 226
二、微观适应性进化的检验方法 / 227
三、宏观适应性进化的检验方法 / 229
四、适应性进化基因 / 232

第三节 分子进化与生物信息学 / 233

一、基因组进化概述 / 233
二、病毒基因组分析 / 234
三、原核生物基因组比较 / 234
四、蛋白质互作网络进化 / 236
五、代谢网络进化分析 / 238

第四节 应用实例：SARS流行病的系统发生分析 / 241

第六章 新一代测序数据分析

CHAPTER 6 NEXT-GENERATION SEQUENCING DATA ANALYSIS

第一节 全基因组测序与重测序 / 246

第二节 新一代测序技术和工作流程 / 248

一、新一代测序法和常见的测序仪 / 249

二、样品准备 / 250

三、合成测序法 / 251

四、第三代测序技术 / 252

第三节 新一代测序数据存储、处理与分析 / 255

一、新一代测序数据格式与质量编码 / 255

二、新一代测序数据库与数据格式转化 / 256

三、测序短片段在参考基因组中的定位 / 256

四、短片段作图软件 / 257

五、基因表达水平估计 / 259

六、可变剪切作图软件包 / 259

第四节 DNA和RNA测序 / 261

一、DNA重测序与个体变异发现 / 261

二、细菌基因组测序与致病性位点发现 / 262

三、宏基因组测序与感染性疾病分析 / 262

四、古生物基因组和进化研究 / 263
五、外显子组测序 / 263
六、非编码RNA测序 / 264
七、核糖体印记与深度测序技术 / 264

第五节　研究实例：基于新一代测序技术的癌症组学研究 / 266

一、短序列数据准备 / 266
二、短序列数据格式转换 / 270
三、短序列在参考基因组上定位 / 272
四、短序列在参考基因组上定位和可变剪切的识别 / 275
五、数字基因表达谱提取 / 279

第七章　转录调控的信息学分析

CHAPTER 7 BIOINFORMATICS ANALYSIS ON TRANSCRIPTION REGULATION

第一节　基因的转录调控 / 284

一、转录 / 284
二、RNA聚合酶 / 285
三、转录调控元件 / 286
四、真核生物转录机制 / 290

第二节　启动子的信息学分析 / 291

一、启动子识别问题 / 291
二、启动子数据资源 / 291
三、真核生物启动子在线分析工具 / 296
四、启动子识别的信息学研究方法 / 297

第三节　转录因子结合位点的信息学研究　/ 300

一、转录因子及其转录调控机制 / 300
二、转录因子结合位点的高通量试验技术 / 303
三、转录因子结合位点相关数据库:
TRANSFAC、JASPAR、SELEX_DB / 306
四、转录因子结合位点模型的建立及分析 / 307
五、利用从头预测算法识别转录因子结合位点 / 308
六、结合Chip-seq等高通量实验数据的
转录因子结合位点预测方法 / 309

第四节　可变剪接的生物信息学研究　/ 312

一、可变剪接的调控机制 / 312
二、可变剪接数据资源: ASD、ASTD / 315
三、利用基因芯片技术进行可变剪接研究 / 315
四、RNA-seq与可变剪接异构体 / 316

第五节　转录调控与人类疾病　/ 318

一、顺式调控元件 / 318
二、反式作用因子 / 319
三、顺式作用元件与疾病（心脏病、肾病、Alzheimer、cancer）/ 320
四、反式作用因子与疾病 / 321

第六节　应用实例　/ 324

一、结合序列和表达谱数据识别组合式
转录调控模式和调控网络 / 324
二、基于启动子结构元件的组合模式识别调控网络 / 324

第八章　生物分子网络与通路

CHAPTER 8 BIOLOGY MOLECULAR NETWORK AND PATHWAY

第一节　网络的基本概念　/ 332

一、网络的定义 / 332

二、有向与无向网络 / 332

三、加权与无权网络 / 333

四、二分网络 / 333

第二节　生物分子网络种类　/ 334

一、蛋白质相互作用网络 / 334

二、基因转录调控网络 / 336

三、代谢和信号传导网络 / 337

四、衍生网络 / 339

第三节　生物分子网络分析方法　/ 342

一、网络的拓扑属性分析 / 342

二、网络的无尺度特性分析 / 345

三、网络的模序搜索 / 346

四、网络的功能模块识别 / 347

五、网络分析软件 / 350

第四节　通路的网络分析方法　/ 353

一、影响分析方法 / 353

二、潜能通路识别分析 / 354

三、通路分析软件 / 356

第五节　应用实例：疾病代谢子通路识别、网络构建和分析 / 358

一、利用通路结构信息识别疾病风险子通路 / 358
二、疾病代谢网络构建和分析 / 359

第九章　复杂疾病的系统遗传学分析

CHAPTER 9 SYSTEM GENETICS ANALYSIS OF COMPLEX DISEASES

第一节　复杂疾病的分子遗传学特征 / 368

一、孟德尔遗传病与复杂疾病 / 368
二、复杂疾病的分子系统特征 / 368
三、重要的复杂疾病数据库 / 369

第二节　变异组信息学与复杂疾病遗传定位 / 375

一、变异组学与人类疾病 / 375
二、SNP与人类复杂疾病定位 / 378
三、变异组学研究资源 / 382

第三节　复杂疾病的分子系统分析 / 389

一、面向通路的基因组范围关联研究 / 389
二、表型性状的分子遗传学 / 391
三、复杂疾病的系统遗传学 / 392

第四节　系统遗传学集成软件工具 / 395

一、Plink软件包与基因互作 / 395
二、基因组范围关联研究软件包SNPtest / 397

三、连锁分析和数量性状分析工具Merlin / 401

第十章 miRNA与复杂疾病

CHAPTER 10 miRNA AND COMPLEX DISEASE

第一节 miRNA与其靶基因 / 408

一、miRNA概述 / 408

二、miRNA靶基因预测 / 410

三、miRNA数据资源 / 412

第二节 miRNA转录组 / 416

一、miRNA表达谱识别癌症相关miRNA / 416

二、miRNA表达谱分类人类癌症 / 417

三、miRNA表达谱与mRNA表达谱整合分析 / 419

四、新一代测序检测miRNA转录组 / 424

第三节 miRNA调控网络 / 428

一、miRNA转录调控网 / 428

二、miRNA功能协同网 / 432

三、miRNA调控不同分子网络 / 438

四、miRNA调控的网络模体 / 443

第四节 miRNA多态和复杂疾病 / 447

一、miRNA基因内部多态 / 447

二、miRNA靶点的多态 / 448

三、miRNA合成机制中的单核苷酸多态 / 450

四、miRNA多态改变表观遗传调控 / 451

第十一章 计算表观遗传学

CHAPTER 11 COMPUTATIONAL EPIGENETICS

第一节 引言 / 456

第二节 表观基因组图谱绘制 / 457

一、绘制基因组范围的DNA甲基化谱 / 457

二、高通量染色质修饰谱的测定 / 461

三、基因组印记与人类疾病 / 466

四、常用的疾病表观遗传学数据库 / 469

第三节 表观遗传修饰谱分析 / 473

一、基因组范围内疾病差异甲基化区域筛选 / 473

二、组蛋白修饰的改变与人类疾病 / 477

三、人类疾病相关的基因组印记分析 / 478

四、常见的疾病表观遗传修饰数据分析软件 / 479

第四节 表观遗传异常标记物识别 / 483

一、DNA甲基化谱的特征在疾病中的应用 / 483

二、整合遗传与表观遗传特征识别疾病相关基因 / 485

第五节 人类表观基因组计划及意义 / 487

一、表观基因组计划介绍 / 487

二、环境表观基因组与人类疾病 / 488

三、人类的表观基因组关联研究 / 488

第六节 应用实例：用于疾病的风险预测、诊断、预后及治疗的表观基因组数据分析 / 490

实例1　人类结肠癌甲基化组差异甲基化分析 / 490
实例2　基于加权蛋白质互作网络优选癌症相关甲基化异常基因 / 493

第十二章　药物生物信息学

CHAPTER 12 PHARMACOBIOINFORMATICS: REVOLUTIONIZING DRUG DISCOVERY RESEARCH

第一节　**药学相关信息资源**　**498**

一、药物综合信息数据库 / 498
二、与药物靶点发现相关数据库 / 499

第二节　**基于生物信息学方法发现潜在药物靶标** / **505**

一、用于药靶发现的生物信息学方法 / 506
二、潜在药靶的生物信息学验证 / 511

第三节　**药物基因组学** / **514**

一、概述 / 514
二、药物基因组学的生物标记 / 514
三、药物基因组学生物标记的预测 / 519
四、药物基因组学生物标记的应用 / 522

中英文名词对照索引 / **525**

绪 论

INTRODUCTION TO BIOINFORMATICS MOLECULAR

第一节

生物信息学的产生及发展

Section 1 The rise and development of bioinformatics

一、生物信息学的产生

生物信息学的产生仅有几十年的时间，bioinformatics这一名词更是在1991年前后才在文献中出现的。事实上，早在1956年，在美国田纳西州盖特林堡召开的首次"生物学中的信息理论研讨会"上，便产生了生物信息学的概念，只不过最初常被称为基因组信息学。就生物信息学的发展而言，它还是一门相当年轻的学科。直到20世纪80~90年代，伴随着计算机科学技术的进步，生物信息学才有了突破性进展。

20世纪后期，生物科学技术、计算机科学技术和网络技术日益渗透到生物科学的各个领域，生物科学的数据资源获得迅猛发展。数据资源的急剧膨胀迫使人们寻求一种强有力的工具去组织这些数据，以利于储存、加工和进一步利用。同时，海量的生物学数据中必然蕴含着重要的生物学规律，这些规律将是解释生命之谜的关键，人们同样需要一种强有力的工具对这些数据进行分析。20世纪80年代末期，生物学家认识到将计算机科学与生物学结合起来的重要意义，开始留意要为这一领域构思一个合适的名称。1987年，"生物信息学"（bioinformatics）这一学科名词诞生。此后，生物信息学的内涵随着研究的深入和现实的需要而几经更迭。1995年，在美国人类基因组计划第一个五年总结报告中，给出了一个较为完整的生物信息学定义：生物信息学是一门交叉科学，它包含了生物信息的获取、加工、存储、分配、分析、解释等在内的所有方面，它综合运用数学、计算机科学和生物学的各种工具，来阐明和理解大量数据所包含的生物学意义。

从生物信息学产生的历程可以看出，基因组信息是生物信息中最早的表现形式，并且基因组信息在生物信息中占有极大的比重。但是，生物信息并不仅限于基因组信息，生物信息学也不等同于基因组信息学。广义地说，生物信息不仅包括基因组信息，如基因的DNA序列、染色体定位，也包括基因产物（蛋白质或RNA）的结构和功能及各生物种间的进化关系等其他信息资源。生物信息学既涉及基因组信息的获取、处理、贮存、传递、分析和解释，又涉及蛋白质组信息学如蛋白质的序列、结构、功能及定位分类、蛋白质连锁图、蛋白质数据库的建立、相关分析软件的开发和应用等方面，还涉及基因与蛋白质的关系如蛋白质编码基因的识别及算法研究、蛋白质结构、功能预测等，另外，新药研制、生物进化也是生物信息学研究的热点。

因此，生物信息学是融合生物科学与数理科学的新兴学科，具体地说生物信息学是以核酸、蛋白质等生物大分子数据库为主要研究对象，以数学、信息学、计算机科学为主要研究手段，以计算机硬件、软件和计算机网络为主要研究工具，对浩如烟海的原始数据进行存储、管理、注释、加工，使之成为具有明确生物意义的生物信息。并通过对生物信息的查询、搜索、比较、分析，从中获取基因编码、基因调控、核酸和蛋白质结构功能及其相互关系等理性知识。在大量信息和知识的基础上，探索生命起源、生物进化以及细胞、器官和个体的发生、发育、病变、衰亡等生命科学中的重大问题。

二、生物信息学的发展历史

生物信息学自产生以来大致经历了前基因组时代、基因组时代和后基因组时代三个发展阶段。三个阶段虽无明显的界限，却真实地反映了生物信息学整个研究重心的转移变化历程。

（一）前基因组时期

从19世纪开始，人们逐渐认识到蛋白质在生命活动中的重要作用。1953年，沃森和克里克发现了DNA双螺旋的结构，开启了分子生物学时代，使遗传的研究深入到分子层次，“生命之谜”被打开，人们清楚地了解遗传信息的构成和传递途径。此后，一些新兴学科如雨后春笋般出现，这些学科的产生和发展为生物信息学的产生奠定了坚实的基础。1956年在美国田纳西州的盖特林堡召开了首次“生物学中的信息理论研讨会”，一些计算生物学家开始进行生物信息相关研究，尽管当时还没有具体地提出生物信息学的概念，但做了许多生物信息搜集和分析方面的工作。1962年，Zucherkandl和Pauling研究了序列变化与进化之间的关系，开创了一个新的领域——分子进化。随后，通过序列比较确定序列的功能及序列分类关系便成为序列分析的主要工作。1967年，Dayhoff研制出蛋白质序列图集，该图集后来演变为著名的蛋白质信息源（protein information resource，PIR）。20世纪60年代是生物信息学形成的萌芽阶段。

从70年代到80年代初期，随着生物化学技术的发展，产生出许多生物分子序列数据，而在这个阶段数学统计方法和计算机技术都得到较快的发展，于是促使一部分计算机科学家应用计算机技术解决生物学问题，特别是与生物分子序列相关的问题。他们开始研究生物分子序列，研究如何根据序列推测结构和功能，出现了一系列著名的序列比较方法，其中，Needleman和Wunsch于1970年提出的序列比对算法是对生物信息学发展最重要的贡献。同年，Gibbs和McIntyre发表的矩阵打点作图法也是进行序列比较的一个著名方法，该方法可用于寻找序列中的重复片段，从而推测其功能。Dayhoff提出的基于点突变模型的PAM（point accepted multation）矩阵是第一个广泛使用的比较氨基酸相似性的打分矩阵，它大大地提高了序列比较算法的性能。1981年，Smith和Waterman提出了著名的公共子序列识别算法，同年，Doolittle提出关于序列模式的概念。1983年，Wilbur和Lipman发表了数据库相似序列搜索算法。1985年，出现快速的蛋白质序列搜索算法FASTP/FASTN，1988年，Pearson和Lipman发表了著名的序列比较算法FASTA。1990年，快速相似序列搜索算法BLAST问世，1997年，BLAST的改进版本PSI-BLAST投入实际应用。

20世纪80年代以后，出现一批生物信息服务机构和生物信息数据库。1982年，核酸数据库GenBank第3版公开发行。1986年，日本核酸序列数据库DDBJ诞生。1986年，出现蛋白质数据库SWISS-PROT。1988年，美国国家卫生研究所和美国国家图书馆成立国家生物技术信息中心NCBI。同年，成立欧洲分子生物学网络（EMBnet），该网络专门发布各种生物数据库。

20世纪90年代后，科学家们开始了大规模的基因组研究。1986年，出现基因组学（genomics）概念，即研究基因组的作图、测序和分析。1990年，国际人类基因组计划启动，该计划被誉为生命科学的“阿波罗登月计划”。1993年，成立Sanger中心，该中心专门从事基因组研究。1995年，第一个细菌基因组被完全测序，1996年，酵母基因组被完全测序。1996年，Affymetrix生产出第一块DNA芯片。1998年，第一个多细胞生物——线虫的基因组被完全测序。1999年，果蝇的基因组被完全测序。1999年年底，国际人类基因组计划联合研究小组宣布人类第一次获得一对完整的人类染色体——第22对染色体的遗传序列。2000年6月24日，人类基因组计划协作组的6个国家研究机构在全球同一时间宣布已完成人类基因组的工作框架图。与此同时，生物信息学在人类基因组计划的推动之下迅速发展。

（二）人类基因组计划

人类基因组计划（human genome project，HGP）是由美国科学家于1985年率先提出，于1990年正式启动的。美国、英国、法国、前西德、日本和中国科学家共同参与了这一预算达30亿美元的人类基因组计划。按照这个计划的设想，在2005年，要把人体内约10万个基因的密码全部解开，同时绘制出人类基因的谱图。换句话说，就是要揭开组成人体4万个基因30亿个碱基对的秘密。人类基因组计划与曼哈顿原子弹计划和阿波罗计划并称为三大科学计划。

人类基因组计划（HGP）的目的是测出人类基因组DNA上30亿个碱基对的序列，发现所有人类基因，找出它们在染色体上的位置，破译人类全部遗传信息。进而解码生命、了解生命的起源、了解生命体生长发育的规律、认识种属之间和个体之间存在差异的起因、认识疾病产生的机制以及长寿与衰老等生命现象、为疾病的诊治提供科学依据。在人类基因组计划中，还包括对五种生物基因组的研究：大肠埃希菌、酵母、线虫、果蝇和小鼠，称之为人类的五种“模式生物”。

人类基因组计划（HGP）的主要任务是人类的DNA测序，包括下面四张谱图，此外还有测序技术、人类基因组序列变异、功能基因组技术、比较基因组学、社会、法律、伦理研究、生物信息学和计算生物学、教育培训等目的，利用HGP发展起来的这些技术和资源进行生物学研究的科学家，促进了人类健康。

1. 遗传图谱（genetic map） 又称连锁图谱（linkage map），它是以具有遗传多态性（在一个遗传位点上具有一个以上的等位基因，在群体中的出现频率皆高于1%）的遗传标记为“路标”，以遗传学距离（在减数分裂事件中两个位点之间进行交换、重组的百分率，1%的重组率称为1cM）为图距的基因组图。遗传图谱的建立为基因识别和完成基因定位创造了条件。意义：6000多个遗传标记已经能够把人的基因组分成6000多个区域，使得连锁分析法可以找到某一致病或表现型基因与某一标记邻近（紧密连锁）的证据，这样可把这一基因定位于这一已知区域，再对基因进行分离和研究。对于疾病而言，找基因和分析基因是关键。

2. 物理图谱（physical map） 物理图谱是指有关构成基因组的全部基因的排列和间距的信息，它是通过对构成基因组的DNA分子进行测定而绘制的。绘制物理图谱的目的是把

有关基因的遗传信息及其在每条染色体上的相对位置线性而系统地排列出来。DNA物理图谱是指DNA链的限制性酶切片段的排列顺序，即酶切片段在DNA链上的定位。因限制性内切酶在DNA链上的切口是以特异序列为基础的，核苷酸序列不同的DNA，经酶切后就会产生不同长度的DNA片段，由此而构成独特的酶切图谱。因此，DNA物理图谱是DNA分子结构的特征之一。DNA是很大的分子，由限制性内切酶产生的用于测序反应的DNA片段只是其中极小部分，这些片段在DNA链中所处的位置关系是应该首先解决的问题，故DNA物理图谱是顺序测定的基础，也可理解为指导DNA测序的蓝图。

3. 序列图谱（sequence map） 随着遗传图谱和物理图谱的完成，测序就成为重中之重的工作。DNA序列分析技术是一个包括制备DNA片段化及碱基分析、DNA信息翻译的多阶段过程。通过测序得到基因组的序列图谱。

4. 基因图谱（gene map） 基因图谱是在识别基因组所包含的蛋白质编码序列的基础上绘制的结合有关基因序列、位置及表达模式等信息的图谱。在人类基因组中鉴别出占2%~5%长度的全部基因的位置、结构与功能，最主要的方法是通过基因的表达产物mRNA反追到染色体的位置。

基因图谱的意义在于它能有效地反映在正常或受控条件下表达的全基因时空图。通过这张图可以了解某一基因在不同时间不同组织、不同水平的表达；也可以了解一种组织中不同时间、不同基因中不同水平的表达，还可以了解某一特定时间、不同组织中的不同基因不同水平的表达。

HGP对人类疾病基因的研究有重要意义，人类疾病相关基因是人类基因组中结构和功能完整性至关重要的信息。对于单基因病，采用"定位克隆"和"定位候选克隆"的全新思路，导致了亨廷顿舞蹈病、遗传性结肠癌和乳腺癌等一大批单基因遗传病致病基因的发现，为这些疾病的基因诊断和基因治疗奠定了基础。对于心血管疾病、肿瘤、糖尿病、神经精神类疾病（老年性痴呆、精神分裂症）、自身免疫性疾病等多基因疾病是目前疾病基因研究的重点。健康相关研究是HGP的重要组成部分，1997年相继提出："肿瘤基因组解剖计划""环境基因组学计划""国际人类基因组单体型图计划（The International HapMap Project）"。

（三）后基因组时代

随着人类基因组计划的完成，我们进入了"后基因组学"（post-genomics）时代。基因组学研究重心已开始从揭示生命的所有遗传信息转移到在分子整体水平对功能的研究上，这种转向的一个标志是产生了功能基因组学（functional genomics）这一新学科。功能基因组学是指在全基因组序列测定的基础上，从整体水平研究基因及其产物在不同时间、空间、条件的结构与功能关系及活动规律的学科。人类基因组计划在基因表达图谱方面已取得一定进展，但它有90%的功能尚不明确，功能基因组学将借助生物信息学的技术平台，利用先进的基因表达技术及庞大的生物功能检测体系，从浩瀚无垠的基因库筛选并确知某一特定基因的功能，通过比较分析基因及其表达的状态，确定基因的功能内涵，揭示生命奥秘，甚至开发出基因产品。功能基因组学在后基因组时代占有重要位置，其研究成果直接给人类健康带来福音。

在后基因组时代生物信息学的作用将更加举足轻重，要读懂人类基因组计划测序得到"天书"，仅仅依靠传统的实验观察手段无济于事，必须借助高性能计算机和高效数据处理的

算法语言。只有如此，"天书"才能发挥它应有的价值。生命科学的革命性巨变已把生物信息学推到了前台，生物信息技术已成为后基因时代的核心技术之一，在蛋白质组学、功能基因组学、药物基因组学等领域必将更有用武之地，从而对生命科学的发展产生无法估计的巨大影响。

三、生物信息学在未来生命科学研究中的作用

21世纪医学模式将发生革命性的变化，生物信息学也将发挥更重要的作用。首先，从19世纪末20世纪初以细胞病理学为基础的医学模式，正在向分子医学（以分子生物学、分子细胞学、分子药理学以及现代计算机技术等为基础）模式转变。人类基因组计划正在建立起人类基因与生理、病理之间关系的知识视图；生物领域的新技术（生物芯片、生物信息学）、新的研究方法（功能基因组学、蛋白组学）在临床中逐步得到应用，更新了医学科学基础。其次，医疗实践以循证医学为主，从基因、蛋白质等大分子水平研究疾病的发病机制，对疾病进行预防、诊断和治疗，目标是向特异性诊断、个体化治疗发展。21世纪，遗传信息在临床环境下的集成应用必将导致个性化医疗等新的临床实践。未来10年预防性基因检测会变得普遍，并将应用在具有家族遗传倾向的个体化监测中，2015年遗传信息将会对临床医学产生普遍影响，医生将通过患者的基因组数据与Internet上可获得的数据库（药物、群体数据、临床档案）进行比较来进行疾病诊断及指导患者治疗；临床医师将能够用计算机输出他们患者的遗传构成，从而能够个性化、有针对性地设计给药。基于遗传信息的决策支持系统、辅助临床医师解释分子标记数据的专家系统、智能化临床决策支持系统等将成为临床医生必不可少的工具。分子水平生物信息检测设备（基因芯片、蛋白质芯片、质谱仪等）将成为医疗领域的新需求。尤其是微流控基因芯片、蛋白质芯片技术将在21世纪成熟并应用于临床，因此生物芯片数据分析技术及分析系统将成为临床医生的常规工具。

此外，伴随着后基因组时代高通量组学（high-throughput omics）技术涌现与生物信息学的飞速发展，出现了大量潜在的生物标记（biomarker），其中一些可以用于疾病诊断和治疗。这些生物标记信息在临床上的应用潜力是巨大的，然而目前仅有少数的标记用于临床实践。如何将这些生物标记应用于临床诊断、疾病风险评估与预防模式、指导个体化治疗、开发新的药物靶点等将是未来生物信息学研究的热点问题，也是转化医学的核心内容。

第二节

生物信息学的主要研究内容

Section 2 The main research content of bioinformatics

生物信息学早期的研究内容主要局限于基因组序列的存储和分析，随着基因组测序数据迅猛增加及计算机技术快速发展，特别是人类基因组计划的顺利完成，产生了海量的生物学数据。这些数据具有丰富的内涵，其中隐藏着丰富的生物学知识。充分利用这些数据，通过数据分析、处理，揭示这些数据的内涵，得到对人类有用的信息，是生物信息学家所面临的一个严峻的挑战。因此，生物信息学的研究内容也在得到不断的丰富和补充。从目前生物信息学的研究内容来看，大致包括以下几个方面：基因组信息学、转录组信息学、蛋白质组信息学和代谢组信息学。

一、基因组信息学

基因组是指一种微生物（包括细菌和病毒）或其他生物体细胞中的总 DNA或RNA（反转录病毒），包括核DNA、细胞器DNA（动植物线粒体DNA和植物叶绿体DNA）和染色体外遗传成分（如细菌的质粒DNA）。随着人类基因组计划（HGP）的实施，产生了大量的基因组信息，分析这些信息是生物信息学的重要内容。人类基因组共有约30亿个碱基对，对如此大量的信息数据进行搜集、存储及分配是生物学领域从未遇到过的问题。这些数据中包括编码人类全部蛋白质和结构核糖核酸（RNA）的信息，以及调控这些蛋白质和核酸装配成生物体的信息。因此解读这些信息是一个很大的难题。基因组信息学的主要目标就是配合人类基因组计划的各项实验研究，测定人类基因组的完整核苷酸序列，确定约10万个人类基因在染色体上的位置，以及研究包括基因在内的各种DNA片段的功能，也就是"读懂"人类基因组。

基因组信息学涉及基因组信息的获取、处理、存储、分配、分析和解释等所有方面。具体而言，就是要构建研究基因组的数据库，发展包括算法、软件、硬件在内的有效的信息分析工具以及完善与基因组研究相关的国际互联网络。随着基因组信息学研究的不断完善和深入，目前生物信息学涉及的基因组信息学研究主要包括比较基因组学、功能基因组学和药物基因组学等。

（一）比较基因组学

比较基因组学（comparative genomics）是基于基因组图谱和测序基础上，对已知的基因

和基因组结构进行比较，来了解基因的功能、表达机制和物种进化的学科。利用模式生物基因组与人类基因组之间编码顺序上和结构上的同源性，克隆人类疾病基因，揭示基因功能和疾病分子机制，阐明物种进化关系，及基因组的内在结构。比较基因组学的基础是相关生物基因组的相似性。两种具有较近共同祖先的生物，它们之间具有种属差别的基因组是由祖先基因组进化而来，两种生物在进化的阶段上越接近，它们的基因组相关性就越高。如果生物之间存在很近的亲缘关系，那么它们的基因组就会表现出同线性(synteny)，即基因序列的部分或全部保守。这样就可以利用模基因组之间编码顺序上和结构上的同源性，通过已知基因组的作图信息定位另外基因组中的基因，从而揭示基因潜在的功能、阐明物种进化关系及基因组的内在结构。

早期的比较基因组研究中，模式生物基因组被用于研究人类疾病基因的功能，利用基因顺序上的同源性克隆人类疾病基因。利用模式生物实验系统上的优越性，在人类基因组研究中的应用比较作图分析复杂性状，加深对基因组结构的认识。此外，通过对不同亲缘关系物种的基因组序列进行比较，能够鉴定出编码序列、非编码调控序列及给定物种独有的序列。而基因组范围之内的序列比对，可以了解不同物种在核苷酸组成、同线性关系和基因顺序方面的异同，进而得到基因分析预测与定位、生物系统发生进化关系等方面的信息。同种群体内的比较基因组研究则发现基因组存在大量的变异和多态性，而正是这种基因组序列的差异构成了不同个体与群体对疾病的易感性和对药物与环境因子不同反应的遗传学基础。目前最常见的变异和多态性包括单核苷酸多态性(single-nucleotide polymorphism, SNP)和拷贝数变异(copy number variant, CNV)。

(二) 功能基因组学

功能基因组学(functional genomics)又被称为后基因组学(post-genomics)，它利用结构基因组所提供的信息和产物，发展和应用新的实验手段，通过在基因组或系统水平上全面分析基因的功能，使得生物学研究从对单一基因或蛋白质的研究转向多个基因或蛋白质同时进行系统的研究。这是在基因组静态的碱基序列弄清楚之后转入对基因组动态的生物学功能学研究。

功能基因组的一个重要任务是进行基因组功能注释(genome annotation)，了解基因的功能，认识基因与疾病的关系，掌握基因的产物及其在生命活动中的作用。在使用全局方法进行研究时，研究人员同时检测大量基因的表达水平，从而在整体水平上获得关于基因功能及基因之间相互作用的信息。如果说生物信息学在人类基因组计划中的着重点是基因组序列的话，那么在功能基因组中，生物信息学的着重点则是序列的生物学意义，基因组编码序列的转录、翻译过程和结果，着重分析基因表达调控信息，分析基因及其产物的功能。在功能基因组时代，应用生物信息学方法，高通量的注释基因组所有编码产物的生物学功能是一个重要的特征。功能基因组学的研究主要包括以下几个方面的内容，并且这几方面都与生物信息学密切相关: ①进一步识别基因，识别基因转录调控信息，分析遗传语言; ②注释所有基因产物的功能，这是目前基因组功能注释的主要层次。序列同源性分析、生物信息关联分析、生物数据挖掘是进行功能注释的主要生物信息学手段; ③研究基因的表达调控机制，研究基因在生物体代谢途径中的地位，分析基因、基因产物之间的相互作用关系，绘制基因调控网络图。

(三) 药物基因组学

药物基因组学(pharmacogenomics)又被称为基因组药物学或基因组药理学,是生物信息学的一个重要分支,定义为在基因组学的基础上,通过将基因表达或单核苷酸的多态性与药物的疗效或毒性联系起来,研究药物如何由于遗传变异而产生不同的作用。药物基因组学根据患者的基因型来保证最大疗效的同时将不良反应降到最低,用于探索合理的方法来优化药物治疗方案。这样的方法使得个体化治疗(personalized medicine)出现,可以根据每个人独特的基因组成来制定最佳的药物或合并用药治疗方案。

药物基因组学可以说是基因功能学与分子药理学的有机结合,在很多方面这种结合是非常必要的。药物基因组学区别于一般意义上的基因学,它不是以发现人体基因组基因为主要目的,而是相对简单地运用已知的基因理论改善患者的治疗。药物基因组学以药物效应及安全性为目标,研究各种基因突变与药效及安全性的关系。正因为药物基因组学是研究基因序列变异及其对药物不同反应的科学,所以它是研究高效、特效药物的重要途径,通过它为患者或者特定人群寻找合适的药物,药物基因组学强调个体化,有重要的理论意义和广阔的应用前景。如当前对基因的研究可发现带有某种特定基因的人,会对某种特定的药物成分,产生某种特定反应。将这个基因、药物成分与服用后反应的一连串关联,运用在用药之上,就可知道带有某特定基因之人,不适合服用含有某特定成分的药物,进而降低药物副作用产生的风险;反之,也可以知道带有某特定基因之人,特别适合服用含有某特定成分的药物,进而提升治愈疾病的几率。

二、转录组信息学

转录组学(transcriptomics)是一门在整体水平上研究细胞中基因转录的情况及转录调控规律的学科。转录组即一个活细胞所能转录出来的所有RNA的总和,是从RNA水平研究基因表达的情况,是研究细胞表型和功能的一个重要手段。转录组是连接基因组遗传信息与生物功能的蛋白质组的纽带,转录水平的调控是最重要也是目前研究最广泛的生物体调控方式。转录组信息学是生物信息学的重要分支,负责研究在特定细胞类型内所生产的RNA分子,探讨在一个特定的细胞群内的基因表达水平和调控情况,通常采用基于DNA芯片技术的高通量技术,最近发展起来的新一代测序技术也广泛用来研究转录组。人类基因组包含有30亿个碱基对,其中大约只有5万个基因转录成mRNA分子,而转录后的mRNA仅部分被翻译生成功能性的蛋白质。与基因组不同,转录组更有时间空间性。我们人体大部分细胞具有一模一样的基因,而即使同一细胞在不同的生长时期及生长环境下,其基因表达情况也是不完全相同的。所以,除了异常的mRNA降解现象(如转录衰减)以外,转录组反映的是特定条件下活跃表达的基因。同时,蛋白质组研究需要更多的转录组研究的信息。因为单一的蛋白质组数据不足以清楚地鉴定基因的功能,因此蛋白质组的数据也需要转录组的研究结果加以印证。因此,转录组的研究可以推断相应未知基因的功能,揭示特定调节基因的作用机制。通过对转录组的研究,科研人员还可以确定不同种类的细胞和组织的基因在何时何地被激活或进入睡眠,对转录本的定量可以了解特定基因的活性和表达量,用于疾病的诊断和治疗,比如与癌症相关的基因表达量的改变可以帮助我们揭开癌症的秘密。

(一) 基因表达图谱

以DNA为模板合成RNA的转录过程是基因表达的第一步，也是基因表达调控的关键环节。所谓基因表达，是指基因携带的遗传信息转变为可辨别的表型的整个过程。与基因组不同的是，转录组的定义中包含了时间和空间的限定。同一细胞在不同的生长时期及生长环境下，其基因表达情况是不完全相同的。通过测序技术揭示造成差异的情况，已是目前最常用的手段。人类基因组包含有30亿个碱基对，其中大约只有5万个基因转录成mRNA分子，转录后的mRNA能被翻译生成蛋白质的也只占整个转录组的40%左右。通常同一种组织表达几乎相同的一套基因以区别于其他组织，如脑组织或心肌组织等分别只表达全部基因中不同的30%而显示出组织的特异性。

转录组谱可以提供什么条件下什么基因表达的信息，并据此推断相应未知基因的功能，揭示特定调节基因的作用机制。通过这种基于基因表达谱的分子标签，不仅可以辨别细胞的表型归属，还可以用于疾病的诊断。同样对那些临床表现不明显或者缺乏诊断金标准的疾病也具有诊断意义，如自闭症。目前对自闭症的诊断要靠长达十多个小时的临床评估才能做出判断。基础研究证实自闭症不是由单一基因引起，而很可能是由一组不稳定的基因造成的一种多基因病变，通过比对正常人群和患者的转录组差异，筛选出与疾病相关的具有诊断意义的特异性表达差异，一旦这种特异的差异表达谱被建立，就可以用于自闭症的诊断，以便能更早地，甚至可以在出现自闭症临床表现之前就对疾病进行诊断，并及早开始干预治疗。转录组的研究应用于临床的另一个例子是可以将表面上看似相同的病症分为多个亚型，尤其是对原发性恶性肿瘤，通过转录组差异表达谱的建立，可以详细描绘出患者的生存期以及对药物的反应等。

(二) 转录调控网络

基因表达是指基因在生物体内的转录、剪接、翻译以及转变成具有生物活性的蛋白质分子之前的所有加工过程。人类基因组大约有两万多个基因，但是在单个细胞中，同时表达的基因往往只有几千甚至几百个，而且很多基因只在特定组织或发育阶段表达。从一套基本不变的基因组中产生出多元化的细胞类型是由调控基因活性的各种信号途径所控制。真核生物转录起始十分复杂，往往需要多种蛋白因子的协助，转录因子与RNA聚合酶Ⅱ形成转录起始复合物，共同参与转录起始的过程。作为基因表达的第一步——转录是调控机制的中心。转录调控因子(transcription factors, TFs)，也称之为反式作用因子(trans-acting factor)有序地结合在目标基因启动子(promoter)序列中的特殊位点，启动基因的转录和控制基因的转录效率。这些位点被称为转录因子结合位点(transcription factor binding sites, TFBSs)，又被称为顺式调控元件(cis-regulatory elements)，其长度从几个到十几个碱基对不等。每个转录因子的结合位点通常都有特定的模式，被称为模体(motif)。找到这些特定的序列片段对研究基因的转录调控有着重要意义。

(三) 转录组测序

基于高通量测序平台的转录组测序技术使能够在单核苷酸水平对任意物种的整体转录活动进行检测，在分析转录本的结构和表达水平的同时，还能够发现未知转录本和稀有转录本，精确的识别可变剪接位点以及cSNP(编码序列单核苷酸多态性)，提供最全面的转录组信

息。相对于传统的芯片杂交平台，转录组测序无需预先针对已知序列设计探针，即可对任意物种的整体转录活动进行检测，提供更精确的数字化信号，更高的检测通量以及更广泛的检测范围，是目前深入研究转录组复杂性的强大工具。

随着二代测序技术的发展，测序成本大幅度降低，大规模转录组测序将成为转录组研究的重要方法。多项研究已经表明，二代测序技术的应用，能有效改善诸如EST、SAGE、CAGE、MPSS、PET和全长cDNA测序等传统转录组研究方法的结果，使之得到大大的提升。基于转录组高通量测序的种种技术优势，此种技术应用范围较广，主要有转录本结构研究（基因边界鉴定、可变剪接研究等），转录本变异研究（如基因融合、编码区SNP研究），非编码区域功能研究（Non-coding RNA研究、microRNA前体研究等），基因表达水平研究以及全新转录本发现。

三、蛋白质组信息学

20世纪90年代中期，在人类基因组计划研究及功能基因组学的基础上，产生了在整体水平上研究细胞内蛋白质的组成及其活动规律的学科——蛋白质组学（proteomics）。蛋白质组学以蛋白质组为研究对象，蛋白质组是某种生物所能表达的所有蛋白质，即包括一种细胞乃至一种生物所表达的全部蛋白质，它们都是由RNA从基因那里转录、剪辑信息后选择性拼接和修饰产生。而RNA转录或RNA剪辑的选择性拼接和转录后的修饰能够产生比基因编码数目多得多的蛋白质，从而成为该种生物巨大的蛋白质组。蛋白质组信息学通过对正常个体及病理个体间的蛋白质组比较分析，找到某些“疾病特异性的蛋白质分子”，它们可成为新药物设计的分子靶点，或者也会为疾病的早期诊断提供分子标志。蛋白质组信息学研究不仅能为生命活动规律提供物质基础，也能为众多种疾病机制的阐明及攻克提供理论根据和解决途径。

（一）结构蛋白质组学

结构蛋白质组学又称组成蛋白质组学，是一种针对有基因组或转录组数据库的生物体或组织、细胞，建立其蛋白质或亚蛋白质组（或蛋白质表达谱）及其蛋白质组连锁群的一种全景式的蛋白组学研究，从而获得对有机体生命活动的全景式认识。大规模的全基因组测序计划正产生越来越多的序列信息，而理解这些信息的关键是理解基因产物——蛋白质的功能。在后基因组时代，蛋白质的三维结构解析是揭示生命密码的重要部分。随着技术进步和大量来自公共机构和私人企业的资金投入，结构蛋白质组学研究开始启动，它的目标是采用工业化生产的方式在基因组规模去大量测定蛋白质的结构。这将会改变结构生物学家的研究方式。蛋白质结构测定的流程，从cDNA的克隆到数据收集，大部分将实现自动化，结构蛋白质组学是实验和理论计算相结合的多学科交叉的领域。目前，结构蛋白质组学仍然面临着许多技术上的挑战，这些挑战也带来了很多机遇，结构蛋白质组学产生的大量结构信息将是一笔巨大的财富，它将给制药行业带来重大变化。近年来，基于蛋白结构的合理药物设计在制药行业非常流行。同时，它也必将给生物学领域带来一场革命。

（二）药物蛋白质组学

将蛋白组学的概念用于药物研究领域，通过对比健康状态与疾病状态的细胞或组织的

蛋白质组表达差异，用于药物研究或药物受体的研究或药物治疗前后蛋白质表达状况的总体，以评价药物类似物的结构与活性关系，寻找高活性的药物，由此发展起来的一门学科称之为药物蛋白质组学。药物蛋白质组学在药物研发过程中起着极其重要的作用，药物蛋白质组学的重要研究内容在临床前包括新药和靶的发现、药物作用模式、毒理学研究，在临床研究方面包括疾病特异性蛋白作为有效患者选择的依据和临床试验的标志。应用类似于药物遗传学的方法，按照蛋白质谱来分类患者，并预测药物作用疗效。蛋白质组学和药物蛋白质组学研究当前还处在一个初期发展阶段，甚至连定义还没有来得及完善，相关的技术手段及其配套应用还很不成熟。但这个领域研究之初，对基础研究和实际应用的期望就表现出强烈结合的趋势。随着蛋白质组学、药物蛋白质组学研究的兴起，人们将在蛋白质水平上重新认识诸如生长、发育和代谢调控等生命活动的规律，为研究重大疾病的机制、疾病诊断、防治和新药开发提供重要的理论基础，并正在成为生物技术药物发展的根本动力，并明显加快新诊断和治疗方法的开发。

四、代谢组信息学

代谢组学（metabonomics or metabolomics）是效仿基因组学和蛋白质组学的研究思想，对生物体内所有代谢物进行定量分析，并寻找代谢物与生理病理变化相对关系的研究方式，是系统生物学的组成部分。其研究对象大都是相对分子质量1000以内的小分子物质。代谢组包括组织细胞代谢组和系统整体代谢组。其中组织细胞代谢组是指是指某个时间点上一个细胞所有代谢物的集合，尤其指在不同代谢过程中充当底物和产物的小分子物质，如脂质、糖和氨基酸等，可以揭示取样时该细胞的生理状态，人类中有上万亿个不同类型的细胞，它们具有潜在不同的组织细胞代谢组。基因和蛋白质只是为细胞发生的活动做准备，活动中大部分实际上是发生在代谢物上，如信号转导、能量转移、细胞间通信都受代谢物调控。进一步说，基因和蛋白表达紧密相连，但代谢物行为更密切地反映出细胞所处的环境，该环境依赖于细胞所摄取的营养状况，所接触的药物和污染物以及其他影响细胞健康的外在因子情况。也可以这么说，基因组学和蛋白质组学只是告诉人们细胞中可能发生的行为，而组织细胞组学告诉人们细胞实际中所发生的行为。而组织细胞代谢组学是研究生物样品，尤其是尿液、唾液和血液中的代谢物谱（主要是指含有哪些代谢物，丰度和分布状况等）变化规律的新学科。

（一）疾病代谢组学

疾病代谢组学作为应用驱动的新兴科学，已在微生物和植物研究、药物毒性和机制研究、疾病诊断和动物模型、基因功能的阐明等领域获得了较广泛的应用，与疾病相关的代谢组学方法与应用研究是目前代谢组学研究的热点之一，广泛应用于病变标志物的发现、疾病的诊断、治疗和预后判断。任何疾病的发生和发展都会影响机体代谢，从而导致体液中代谢物质发生显著变化，通过比较机体生理与疾病状态，甚至是同一疾病不同分型的代谢物的不同，将能找到与疾病诊断及分型相关的标志性代谢物，从而发现表征这些疾病的化学特征模式，代谢组学正好适应这一发展趋势。

生物机体的代谢在正常情况下处于一种动态的平衡中，而当机体患病或出现某种病变，

就会打破这种动态的平衡，引起机体内部代谢的紊乱，而这些代谢的紊乱，也通常会使机体的血液、尿液或其他组织液发生一定的变化。因为机体的正常生理活动需要通过体内的各个循环系统的平衡协作而得到保证，包括血液循环、尿的排泄。对尿液和血液等体液代谢组进行检测和分析，就有可能对疾病从发病到病情不断变化的整个过程进行了解和认识，就有可能发现与疾病发生相关的生物标志物并认识相关的病理发生机制，就可以对疾病在其发病之前或发病之初进行预防、诊断和治疗，或者根据疾病不同阶段的特征进行个性化的治疗，达到更好的治疗效果。

（二）药物代谢组学

药物代谢组学（pharmacometabonomics）是研究药物作用于细胞靶分子之后所形成的代谢产物的分子特征的科学。从人类组织及体液，如汗液、血液、尿液等这些人类生命过程代谢物质中药物作用过程中的代谢物分子的分析可以推断药物作用于靶分子的过程，用于阐述药物作用的化学机制。不同于传统的药物代谢动力学，药物代谢组学不仅仅关注药物分子本身在作用于靶分子后的代谢产物，还关注药物与靶分子和非靶分子作用后的代谢产物，以及这些产物之间以及它们与无药物作用的代谢产物发生化学反应之后的产物。

第三节

当前生物信息学研究的热点

Section 3 The hotspot of current bioinformatics research

自从1987年出现bioinformatics这一词汇以来，生物信息学的研究任务随着科研和现实需要的变化而几经更迭。当前，一般认为，生物信息学主要是一门研究生物学系统和生物学过程中的信息流的综合系统科学，通过它独特的桥梁作用和整合作用，使我们能够从各生物学科中众多分散的观测资料中获得对生物学系统和生物学过程的运作机制的理解，最终达到自由应用于相关实践的目的。例如，就疾病而言，生物信息学就是要系统地理解导致机体功能异常的生物机制并从而得出科学的治疗方案；就生物演化而言，生物信息学就是要系统地解释生物界演化的从微观分子水平到宏观形体功能水平的根本原则，从而使人类更好地认识自己在自然界中的地位，科学地认识和改造人类的未来。因此与以往相比，生物信息学无论从认识水平上还是从实践水平上都开创了一种崭新的模式。

一、新一代测序数据的生物信息学分析

DNA测序（DNA sequencing）作为一种重要的实验技术，在生物学研究中有着广泛的应用。早在DNA双螺旋结构（Watson and Crick，1953）被发现后不久就有人报道过DNA测序技术，但是当时的操作流程复杂，没能形成规模。随后在1977年Sanger发明了具有里程碑意义的末端终止测序法，同年A.M.Maxam和W.Gilbert发明了化学降解法。Sanger法因为既简便又快速，并经过后续的不断改良，成为迄今为止DNA测序的主流。然而随着科学的发展，传统的Sanger测序已经不能完全满足研究的需要，对模式生物进行基因组重测序以及对一些非模式生物的基因组测序，都需要费用更低、通量更高、速度更快的测序技术，新一代测序技术（next-generation sequencing）应运而生。新一代测序技术的核心思想是边合成边测序（sequencing by synthesis），即通过捕捉新合成的末端标记来确定DNA的序列，现有的技术平台主要包括Roche/454 FLX、Illumina/Solexa Genome Analyzer和Applied Biosystems SOLID System。

随着高通量新一代测序技术的快速发展，DNA测序（DNA-seq）、RNA测序（RNA-seq）已成为基因组、转录组分析的新的重要手段，也为生物信息学研究开创了崭新的局面。新一代测序可一次性获得数百万甚至数十亿的序列数据信息，开发能够快速鉴定出不同组织、不同发育阶段、不同疾病状态下的转录本及其表达差异的生物信息学理论和方法，为基于新一代测序技术的复杂疾病研究提供有力工具，是当前生物信息学研究的重要任务之一。

二、非编码区序列分析与功能识别

非编码DNA(或称“垃圾DNA”),是指不包含制造蛋白质的指令,或是只能制造出无转译能力RNA的DNA序列。此类DNA在真核生物的基因组中占大多数。有很长一段时间科学家们没有认识到这些非编码的作用,因此,这些重复的DNA片段被冠以垃圾DNA的称号。随着时间推移,科学家们对垃圾DNA的认识逐渐深入,慢慢地发现其实很多非编码DNA有着其独特的作用,它们在基因剪切等方面起重要的作用。

科学家们已经发现:“垃圾”DNA的功能之一就是调节基因的活动,如同一道指令一样,控制着基因。一些控制基因开和关的特殊蛋白(转录因子)能特异识别基因附近的非编码“垃圾”DNA,通过与它们相互作用参与基因的抑制与激活。科学家还发现,大多数基因的开启和关闭是由附近的“垃圾”DNA控制的。它们就像是基因的“分子”开关,调节基因的活动。许多“垃圾”DNA序列的变化与复杂疾病如关节炎、共济失调症等的发生息息相关。不同个体对药物的反应、对疾病易感性的差异在很多情况下也是由一些特殊的“垃圾”DNA调节的。甚至一些科学家猜想:可能正是“垃圾”DNA造成了人类个体间的差异。迄今为止,细胞中的rRNA、tRNA、snRNA、asRNA、snoRNA、miRNA、piRNA都非编码“垃圾”DNA合成。它们参与到基因活化、基因沉默、基因印记、剂量补偿、蛋白合成与功能调节、代谢调控等众多生物学过程中。

在过去十年里,与复杂疾病关联的微小RNA(microRNA, miRNA)的研究取得了不少成果。miRNA是一类非编码的小RNA分子,其长度约22个核苷酸(nucleotide,简称nt),通过和其靶基因3’非翻译区(3’ untranslated region,简称3’ UTR)结合引导RNA诱导的沉默复合体(RNA-induced silencing complex,简称RISC)促进其靶mRNA的降解或阻碍其靶mRNA的翻译。大量研究表明miRNA可以通过精细地调节基因的转录表达进而参与细胞的发育、分化、增殖、凋亡以及应激反应等生物学过程。研究人员发现其在复杂疾病的发生发展过程中起着巨大的作用,其功能异常能够导致各种人类复杂疾病(如癌症、心血管疾病等)的发生,这使miRNA成为疾病诊断、预后的新的生物学标记(biomarker),并为进一步揭示复杂疾病的发病机制提供了新的方向。随着对复杂疾病关联的非编码RNA研究的深入,近年来的研究逐渐转向长链非编码RNA(long noncoding RNA, lncRNA)。lncRNA是一类转录本长度超过200nt的RNA分子,它们并不编码蛋白,而是以RNA的形式在多种层面上调控基因的表达水平,如表观遗传调控、转录调控和调控蛋白活性,改变RNA的剪切模式以及转录后调控等。目前研究所展现出的lncRNA繁多的分子生物学功能,为人们研究调控领域提出了崭新的视角。lncRNA通过与DNA、RNA、蛋白质的相互作用,在生命活动调控网络中扮演着十分重要的角色。除了在基因表达调控方面发挥着十分重要的作用,lncRNA与物种进化、胚胎发育、物质代谢以及复杂疾病的发生等都有着紧密的联系。

三、整合信息组学

当前,由各种“omics”组学技术,如基因组学、转录组学、蛋白质组学和代谢组学等技术,积累了大量的实验数据。我们面临的挑战是如何从这些组学数据中,利用已有的生物信息

学的技术手段，在新的系统层次、多水平、多途径来了解生命过程。鉴于此，人们希望形成一个生物信息学的特定领域，以便解决这些很重要的问题，这就是"整合信息组学"。

用系统生物学的观点，整合各类"omics"组学信息，发展系统整合语言，提出细胞与组织乃至人体的生理和病理的数字化模型，运用系统整合语言发展与中心法则有关的模型与假说，并在实验和临床中加以验证，提出药物与靶点相互作用及其网络作用的模型与假说，并在实验和临床中进行验证，为重大疾病的防治、诊治提供理论依据。随着基因组研究的完成，以及向功能基因组研究的转化，将基因组、转录组、蛋白质组以及比较基因组学的数据综合集成，构建基因调控网络，从系统的角度来研究生物学，为系统生物学的研究提供工具，成为生物信息学的研究重点。此外，新一代测序等高通量技术的应用，产生海量的基因表达数据，这些数据中隐含了基因表达控制的信息，对这些的分析和挖掘，以及数据的标准化已成为生物信息学的研究热点。

四、转化医学和临床生物信息学

转化医学（translational medicine），又被称作转化研究（translational research），是近年来国际医学科学领域出现的新概念，是基因组和生物信息学革命的时代产物，通过研究可诊断及监测人类疾病的新参数——生物标志物，为开发新药品、新诊断方法、新治疗方法开辟出一条具有革命性意义的新途径。转化医学研究的主要任务是，将基础研究所取得的成果尽快转化为临床问题的解决方法；将基础研究获得的知识、成果快速转化为临床上的治疗新方法，以及把临床医疗的实际情况反馈给实验室并以此来完善相关课题的基础研究并进一步开展新的研究的一种双向过程，即"从实验室到病房（bench to bedside）"和"从病房到实验室（bedside to bench）"双向通道研究，简称为B2B。

临床生物信息学的目的是应用生物信息学知识和技术来帮助诊断、治疗、预防和控制疾病，以及发展化学的、结构的和生化的方法来应用于临床研究。癌症研究中，在癌症发生的不同阶段，如起始、持续和发展时期，生物信息学工具被用于检测几种癌症的生物标记。根据NCI的解释：生物标记的定义是细胞的、生化的、分子的（遗传和表观遗传）改变。有了生物标记，一个正常的、异常的或简单的生物学过程就可以被识别或监测。生物标记可以通过生物媒介，如组织、细胞或流体来衡量，也可用于评估癌症的早期诊断、风险、癌症分类和预断癌症病情。

五、生物信息学与新药研究

当前生物信息学的一个重要任务是辅助药物设计和新药研发。新药研究和开发是一项耗资巨大的工程。过去，每一种新药从研发到投入市场平均需要10~15年，耗费数十亿美元。而现在，生物信息技术为药物研究设计提供了崭新的研究思路和手段，生物信息学所提供的数据和软件可以指导对药物作用靶位的选定和药物分子的设计。这种方法有快速、高效的特点，它的研究范围包括大分子结构功能的模拟和预报、药物分子与大分子结合的模拟、生物分子在指定细胞的分布和位点等。生物信息学已经在新药设计的各个环节，如初始阶段、筛选及药物设计，以及新药开发阶段发挥着越来越重要的作用。利用强大的计算工具，新药

开发平均费用时间都大大降低了。

传统药物研究中，可供筛选的化合物数量有限，新药发现的速度很慢，耗资巨大，成功率也很低。生物信息学在筛选及药物设计中的应用，给药物发现带来了新的机遇。在"人类基因组计划"完成后，药物筛选有了很大发展。主要是运用计算机技术，以药物靶标分子三维结构和蛋白质晶体结构为基础，对含有大量化合物结构的数据库进行模拟"筛选"，迅速高效地发现先导化合物及其新用途。这种药物设计的方法是根据靶标分子与药物分子相结合的活性部位的几何形状和化学特征，设计出与其相匹配的具有新颖结构的药物分子。使用这种方法需建立大量化合物的三维结构数据库，然后将库中的分子分别与靶标分子结合，通过不断优化小分子化合物的位置以及分子内部柔性键的二面角，寻找小分子化合物与靶标大分子作用的最佳构象，计算其相互作用及结合能。在库中所有分子均完成特异结合计算之后，即可以从中找出与靶标分子结合的最佳分子。

生物信息学不仅有助于药物靶基的发现、药物设计与药物筛选，而且还有利于药物开发的临床研究。这主要表现在单核苷酸多态（SNP）、药物基因学（pharmacogenomics）和药物遗传学的研究及结果的应用。例如，通过SNP与药物反应的相关分析能够显示出在不同个体的药物作用目标或药物代谢途径中存在某个酶的差异，揭示个体的基因组多态与疾病治疗药物反应之间的关系。这就让我们可以预测出哪种药或疫苗对哪些携带特殊基因型的个人最有效，因此医生就可以根据不同患者对药物的不同反应，进行个体化给药与个体化治疗，提高治疗效果，增加临床试验的成功率，促进个体化药物的开发。

综上，复杂疾病的治疗，逐渐走出实验室，迅速进入转化研究阶段，其重要标志，就是依据基因组学或蛋白组学的临床研究。复杂疾病的发生与发展是一个多基因参与、多步骤、复杂的生物学过程，仅仅依据病理类型、临床分期以及患者年龄、行为状态等临床特征选择治疗方法以远远达不到个体化数字化治疗的要求。通过生物信息学的方法研究复杂疾病的组学谱，全面详尽地了解肿瘤的生物学特性来指导临床治疗，是未来医疗的必由之路。

毋庸置疑，以DNA和蛋白质序列为源头的生物信息学，已经显著改变了传统实验数据的处理手段，变革了基础生命科学的运作方式，推进了应用生物技术及相关学科的发展速度。随着生物信息学研究的不断深入和扩展，势必带来整个生物领域的重大革命，尤其对人类基因疾病的诊断和治疗以及药物开发必将产生深远影响。

（李　霞）

第一章
CHAPTER 1

序列比对与序列特征分析
SEQUENCE ALIGNMENT AND ANALYSIS OF SEQUENCE CHARACTERISTICS

随着近年来生物实验技术和方法的快速发展，通过实验获取的RNA、DNA和蛋白质序列数据以前所未有的速度增长。世界各国的生物学家和计算机学家合作通过对这些序列数据的分类、收集和整理构建了基因组数据库、核酸和蛋白质一级结构序列数据库以及在此基础上构建特殊类型的核酸和蛋白质序列数据库。对各种生物序列进行分析是生物信息学最主要的研究内容之一，它可以分为两个主要部分：一是序列之间的比较分析。二是序列组成和特征分析。序列比较的基本操作是比对，将未知序列同已知序列进行相似性比较是一种强有力的研究手段，从序列的片段测定、拼接、基因的表达分析，到RNA和蛋白质的结构功能预测，物种亲缘树的构建都需要进行生物分子序列的相似性比较。生物信息学中的序列比对算法的研究具有非常重要的理论意义和实践意义。而对DNA序列和蛋白质序列进行序列特征分析，能够从分子层面上解读基因的结构特点，了解与基因表达调控相关的信息，明确DNA序列与蛋白质序列之间的编码关系，为进一步揭示基因的结构和功能，研究蛋白质结构和功能之间的关系提供理论依据。

第一节

获取DNA、RNA和蛋白质序列

Section 1　DNA, RNA and Protein Sequence Information Resources

一、DNA序列的获取

（一）国际核酸序列数据库协会

1988年由国际上三大主要的公共核酸序列数据库共同建立了国际核酸序列数据库协会（international nucleotide sequence database collaboration，INSDC，http://www.insdc.org/）（图1–1），这三个数据库分别是位于美国马里兰州的贝塞斯达的美国国家生物技术信息中心（NCBI）的GenBank，位于英国的欧洲分子生物学研究中心（EMBL）的ENA和日本的DNA数据库（DDBJ）。三大核酸数据库之间各自搜集世界各国有关实验室和测序机构所发布的序列数据，每天将新测定或更新的数据进行交换，实现了全球范围内核酸序列的同步更新和交换共享。

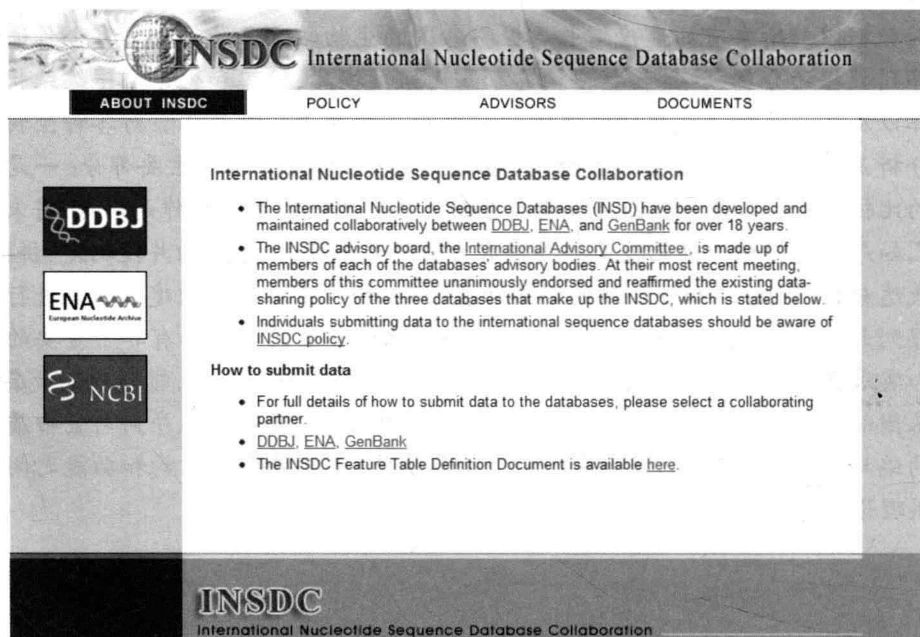


图1–1　INSDC的主页

1. GenBank　GenBank（http: //www.ncbi.nlm.nih.gov/genbank/）是由隶属于美国国立卫生院（national institute of health，NIH）的美国国立生物技术信息中心（national center for biotechnology information，NCBI）建立的国际权威核酸序列数据库。NCBI构建的GenBank数据库序列来自于发现者提交的序列、批量提交的表达序列标签（expressed sequence tag，EST）、基因组测序序列（genome survey sequences，GSS）和其他测序中心提交的高通量数据以及美国专利商标局提供的已发表的专利序列数据。截止到2010年，GenBank 共收录了超过38万个物种的198 156 212条序列，总长度超过了3000多亿个碱基。图1-2总结了从1982~2008年GenBank中DNA序列和碱基数目的变化情况。除了序列信息以外，GenBank还收录了相应的参考文献记录和生物学注释。

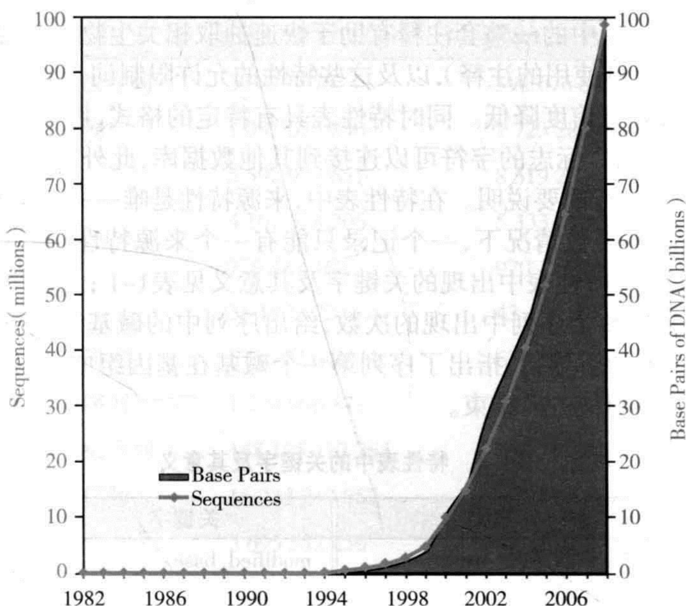


图1-2　GenBank中DNA序列和碱基数量的变化情况

来源于：http: //www.ncbi.nlm.nih.gov/genbank/genbankstats.html

（1）GenBank数据库的组织结构

1）GenBank数据库中的序列文件和序列条目

完整的GenBank数据库包括序列文件，索引文件以及其他有关文件。索引文件是根据数据库中作者、参考文献等子段建立的，用于数据库查询。GenBank中最常用的是序列文件。序列文件的基本单位是序列条目，包括核苷酸碱基排列顺序和注释两部分。序列文件由单个序列条目组成，每个条目是一个纯文本文件，序列条目由字段组成，每个字段由关键字起始，后面为该字段的具体说明。有些字段又分若干子字段，以次关键字或特性表说明符开始。每个序列条目以双斜杠“//”作结束标记。序列条目的格式非常重要，关键字从第一列开始，次关键字从第三列开始，特性表说明符从第五列开始。每个字段可以占一行，也可以占若干行。若一行中写不下时，继续行以空格开始。每条GenBank序列条目的关键字包括代码（LOCUS），说明（DEFINITION），编号（ACCESSION），标识符（NID），关键词（KEYWORDS），数据来源（SOURCE），文献（REFERENCE），特性表（FEATURES），碱基组成（BASE COUNT）和排列顺序（ORIGIN）。

代码行是该序列条目的标记，或者说标识符，蕴涵这个序列的功能。其中，检索号是唯一的、不可重复的。该字段还包括其他相关内容如序列长度、类型、众数以及录入日期等；说明字段是关于这一序列的简单描述，用以总结记录的生物学意义；编号字段具有唯一性和永久性，在文献中引用这个序列时，应该以此编号为准；核酸标识符提供了序列信息的当前版本；关键词字段由该序列的提交者提供，包括该序列的基因产物以及其他相关信息；数据来源字段说明该序列是从何种生物体、何种组织得到的；次关键字种属（ORGANISM）指出该生物体的分类学地位，如人、真核生物等；文献字段说明该序列中的相关文献、包括作者、题目及期刊名称等，以次关键词列出。该字段中还列出医学文献摘要数据库MEDLINE的代码。该代码实际上是个网络链接指针，点击它可以直接调用上述文献摘要。一个序列可以有多篇文献，以不同序号表示，并给出该序列中的哪一部分与文献有关；特性表直接给出了记录的生物学背景知识，记录中的一整套注释有助于快速抽取相关生物学信息。特性表详细地描述了合法的特性（允许使用的注释），以及这些特性的允许限制词，如果这些注释仅仅是推测或是计算得到的，其可信度降低。同时特性表具有特定的格式，用来详细描述序列特性。特性表中带有‘/db-xref/’标志的字符可以连接到其他数据库，此外还对翻译所得的信号肽以及最终蛋白质产物进行简要说明。在特性表中，来源特性是唯一一个必须在所有GenBank记录中出现的特性，大多数情况下，一个记录只能有一个来源特性，并带有‘/organism’限定词，在GenBank注释的特性表中出现的关键字及其意义见表1-1；碱基组成是碱基含量字段，计算出不同碱基在整个序列中出现的次数，给出序列中的碱基组成；GenBank数据库记录以ORIGIN行为序列的引导行，指出了序列第一个碱基在基因组中的可能位置，最后列出全部的碱基序列，以双斜杠“//”结束。

表1-1 特性表中的关键字及其意义

关键字	意义	关键字	意义
3’ UTR	3’ 非翻译区	modified_base	修饰过的碱基
5’ UTR	5’ 非翻译区	mRNA	信使RNA
-10_signal	-10 信号	mutation	突变
-35_signal	-35 信号	rRNA	核糖体RNA
CAAT_signal	CAAT信号	tRNA	转运RNA
CDS	编码序列，含终止密码子	polyA_signal	多聚A信号
enhancer	增强子	polyA_site	多聚A位点
exon	外显子	prim_transcript	初始转录码
GC_signal	GC信号	promotor	启动子
gene	已命名的基因序列	protein_bind	蛋白质结合位点
intron	内含子	rep_origin	复制起点
LTR	长终端重复序列	repeat_region	重复区
mat_peptide	翻译后被修饰的序列，不含终止密码子	repeat_unit	重复单元
mis_binding	错结合点	satellite	卫星片段
misc_feature	其他性状	sig_peptide	信号肽
misc_RNA	其他RNA	TATA_signal	TATA 信号
mis_signal	其他信号	terminator	终端子

2）GenBank数据库中的子库

GenBank数据库中的序列记录可以划分为11个子库（BCT，INV，MAM，PHG，PLN，PRI，ROD，SYN，UNA，VRL，VRT）和7个高通量子库（ENV，EST，GSS，HTC，HTG，STS，TSA）。子库的划分可以把数据库查询限定在某一特定部分加快了查询速度。同时，基因组计划快速测序得到的大量序列尚未加以注释，将它们单独分类，有利于数据库查询和搜索。表1-2中显示了这些子库中碱基的数目和增长趋势。同时，GenBank数据库中全基因组测序数据也在不断的增加，现在已经有超过1200种细菌和古细菌及460多种脊椎动物的全基因组拼接数据。

表1-2 GenBank子库中碱基的数目和增长趋势

子库名称	描述	版本173（8/2009）	版本179（8/2010）	增长速率（%）
TSA	转录组鸟枪法序列	39 829 979	398 676 845	900.9
ENV	环境样本序列	1 091 072 890	1 723 286 428	57.9
PAT	专利序列	5 592 927 651	8 519 294 473	52.3
BCT	细菌序列	4 107 328 206	5 333 010 385	29.8
VRL	病毒序列	779 481 462	970 125 245	24.5
PHG	噬菌体序列	36 100 172	43 456 808	20.4
MAM	其他哺乳类序列	576 977 646	679 274 390	17.7
INV	非脊椎动物序列	1 734 996 371	2 036 240 836	17.4
WGS	全基因组鸟枪序列	148 165 117 763	169 253 846 128	14.2
GSS	基因组测序序列	16 738 219 857	18 442 479 673	10.2
PLN	植物序列	3 695 552 256	4 038 424 961	9.3
SYN	人工合成序列	131 361 806	142 548 355	8.5
VRT	其他脊椎动物序列	2 366 300 257	2 533 789 261	7.1
EST	EST序列	34 522 977 161	36 803 930 321	6.6
HTC	高通量cDNA序列	636 472 189	659 355 057	3.6
PRI	灵长类序列	5 751 413 009	5 943 029 356	3.3
ROD	啮齿类序列	4 206 718 960	4 298 354 944	2.2
HTG	高通量基因组序列	23 895 733 886	24 276 862 305	1.6
UNA	未经注释的序列	119 348	120 289	0.8
STS	序列标签位点	629 573 650	634 263 196	0.7
Total	GenBank中的序列	254 698 274 519	286 730 369 256	12.6

注：来源于Benson DA，Karsch-Mizrachi I，Lipman DJ，et al. GenBank. Nucleic Acids Res，2001；39：32-37.

转录组鸟枪法组装序列（TSA）子库是伴随着新一代测序技术出现的，如"Roche-454 Life Science"、"Illumina Solexa"和"Applied Biosystems SOLID"，新加入到GenBank数据库中的一个新的子库。该子库中的序列主要由来自于NCBI的踪迹档案（trace archive，TA）序列读取档案（sequence read archive，SRA）和EST数据子库的序列组装而成。

环境样本序列(ENV)数据子库中的序列来自于非WGS测序的序列,这些序列的物种来源是未知的。许多的环境样本序列来自于不同动物组织(如内脏、皮肤)或者淡水沉积物、温泉和矿井污水区等特殊环境中的微生物。环境样品序列记录中在关键字字段标明“ENV”并且在来源特征处标明“/environmental_sample”。

表达序列标签(EST)子库一直是研究基因表达及基因注释的重要的资源,同时也是非WGS子库中最大的一个。它收录了一系列物种的“测序一次”的cDNA序列或者是表达序列标签。截止到2011年1月,EST数据库(100111版本)共收录了7.09千万条记录,表1-3列出了收录最多的前10个物种。EST数据库中的数据可以通过NCBI的FTP站点免费下载ftp.ncbi.nih.gov/repository/dbEST。EST数据库中的数据经过进一步blast程序的同源比对生成了UniGene数据库(www.ncbi.nlm.nih.gov/unigene)。UniGene数据库中已经存储了120个物种的430万个簇条目。

表1-3 EST数据库中记录数量居前的10个物种(dbEST第100111版本,2011年1月)

物种	通用名	EST记录数
Homo sapiens	人	8 315 272
Mus musculus + domesticus	小鼠	4 853 562
Zea mays	玉米	2 019 114
Sus scrofa	野猪	1 624 046
Bos taurus	牛	1 559 494
Arabidopsis thaliana	拟南芥	1 529 700
Danio rerio	斑马鱼	1 488 275
Glycine max	大豆	1 461 624
Xenopus (Silurana) tropicalis	热带爪蟾	1 271 375
Oryza sativa	水稻	1 252 989

全基因组鸟枪序列(whole genome shotgun, WGS)数据字库是WGS重叠拼接组序列,每条序列都有一个访问号,该访问号包含一个4字母的计划ID号,后面是两个数字的版本号和六个数字的重叠拼接的ID号。如果一个WGS计划的访问号是“XXXX00000000”,那么这个计划的第一个组装版本是XXXX01000000,第一个重叠群的版本是XXXX01000001。截止到2010年10月,全基因组鸟枪测序计划已经向GenBank数据库提交了6.4千万条拼接序列,构建了800万个大规模的染色体骨架的组装体。

高通量基因组(HTG)和高通量cDNA(HTC)序列子库:HTG子库(www.ncbi.nlm.nih.gov/HTGS)是GenBank数据库中一个存储尚未完成的大规模基因组记录的数据子库。这些记录可以根据数据质量分为0~3个阶段,3阶段代表完成状态。一旦达到3状态,HTG中的记录就会被转移到合适的GenBank数据库的其他子库中。GenBank中的HTC子库存储高通量的cDNA序列,这些序列是一些初级序列,可能包含3’和5’端的非翻译区、部分编码区和内含子。完成后的高质量的HTC序列也将会转移到合适的GenBank数据库的其他子库中。

3)基于物种的分类

GenBank数据库中的序列还可以根据物种名进行检索。表1-4总结了在GenBank数据库

中20种非鸟枪法测序（non-WGS）最多的物种和碱基数目。

表1-4 GenBank数据库中碱基数量居前的20个物种

物种	通用名	碱基/个
Homo sapiens	人	14 792 487 417
Mus musculus	小鼠	8 859 010 528
Rattus norvegicus	大鼠	6 443 768 086
Bos taurus	牛	5 361 712 195
Zea mays	玉米	5 037 629 354
Sus scrofa	野猪	4 783 381 701
Danio rerio	斑马鱼	3 137 945 523
Strongylocentrotus purpuratus	紫海胆	1 352 920 226
Oryza sativa Japonica Group	水稻	1 197 245 122
Nicotiana tabacum	烟草	1 187 388 273
Xenopus（Silurana）tropicalis	热带爪蟾	1 147 132 278
Drosophila melanogaster	果蝇	1 047 707 620
Pan troglodytes	黑猩猩	1 001 926 471
Arabidopsis thaliana	拟南芥	1 001 073 627
Canis lupus familiaris	家犬	943 043 649
Vitis vinifera	葡萄	913 911 649
Gallus gallus	鸡	891 463 513
Glycine max	大豆	886 103 518
Macaca mulatta	猕猴	821 393 285
Ciona intestinalis	海鞘	748 350 657

注：来源于Benson DA, Karsch-Mizrachi I, Lipman DJ, et al. GenBank. Nucleic Acids Res,2011;39：32-37.

（2）在GenBank数据中获取核酸序列的方法

1）Entrez检索系统

Entrez（http://www.ncbi.nlm.nih.gov/sites）是NCBI的数据库检索查询系统。利用Entrez系统用户可以方便地检索GenBank数据库中的核酸序列。GenBank数据库中的EST子库和GSS子库就存储在Entrez的EST和GSS数据库中，GenBank数据库中其他的记录存储在Entrez的Nucleotide数据库中。用户可以利用Entrez界面上提供的限制条件（Limits）、索引（Index）、检索历史（History）和剪贴板（Clipboard）等功能来实现复杂的检索查询工作。对于检索获得的记录，用户可以选择需要显示的数据，保存查询结果，甚至以图形方式观看检索获得的序列。更详细的Entrez使用说明可以在该主页上获得。用户利用Entrez系统还可以检索GenBank和其他资源的蛋白质序列、基因组图谱、基因表达数据、NCBI分类数据和蛋白质结构数据，以及PubMed和PubMed Central 中的学术文献。

2）与测序计划相关的序列记录

由NCBI建立的基因组计划数据库允许测序中心登记测序计划，并且得到一个特有的计划标识符，用以保证测序计划和该计划产生的数据之间建立可靠的联系。该数据库后来被重新命名为BioProjects（http: //www.ncbi.nlm.nih.gov/bioproject）。BioProjects数据库通过一种有组织的结构允许用户访问各种研究计划及其产生的数据。

3）BLAST序列相似性搜索

序列相似性搜索是GenBank数据库中数据注释的最基础和使用最多的方法。NCBI提供了一系列的BLAST程序用于检测查询序列和数据库中序列的相似性。用户可以在NCBI网站上提交一段序列与NCBI的数据库中的序列进行相似性比对，也可以通过NCBI的FTP下载本地BLAST软件后，在本地做BLAST相似性比对。

4）通过GenBank的FTP站点

NCBI以传统的文本文件格式发布GenBank的数据，并且以ASN.1格式进行内部维护。每两个月的GenBank以及EMBL和DDBJ的更新的序列数据都可以从NCBI的匿名FTP服务器上下载（ftp.ncbi.nih.gov/genbank）。同时还可以从NCBI的FTP服务器上免费下载完整的库。在GenBank发布的第179版本中有1443个文件，需要大概484GB的存储空间。

2. ENA数据库

European nucleotide archive（ENA）是欧洲的主要的核酸序列数据库，由欧洲分子生物学研究中心（European molecular biology laboratory, EMBL）的欧洲生物信息学研究所（European bioinformatics institute, EBI）建立和维护（图1–3）。ENA数据库整合了原始的序列数据、组装信息和功能注释。ENA数据库主要包括三个主要的数据库：序列读取数据库（sequence read archive, SRA），测序数据库（trace archive）和EMBL数据库（EMBL–Bank）。ENA的目标是通过提供数据提交、存储、搜索和下载服务支持和促进核酸测序的发展。截止到2010年10月ENA数据库一共存储了5000亿个原始和组装的序列，包括50兆的碱基。在最近三年，在SRA中存储的新一代测序技术产生的序列已经成为ENA中最大和增长最快的数据，已经占到了ENA数据大约95%。同时，在ENA中也存储了超过1400种单细胞和多细胞生物和3000多种病毒和噬菌体的全基因组序列。

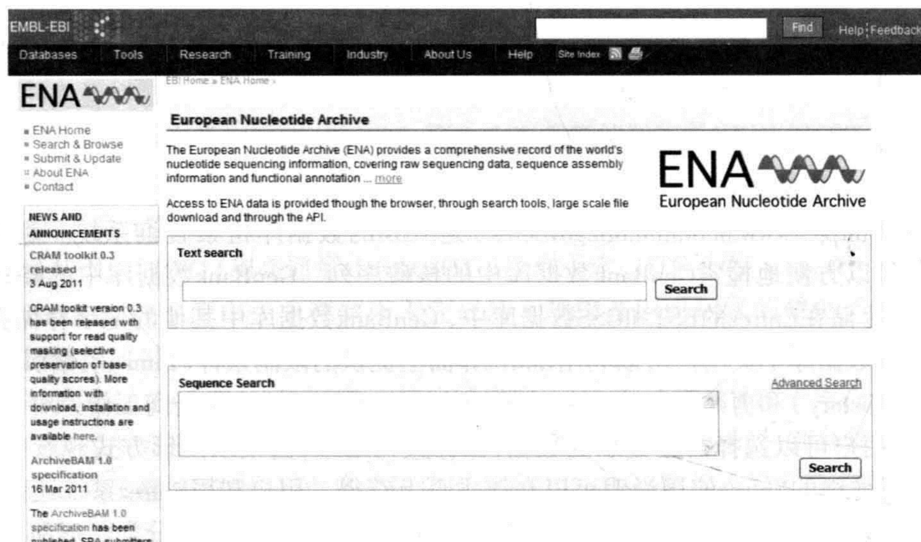


图1–3 ENA数据库的主页

ENA的数据可以通过网络浏览器以XML、HTML、FASTA和文本格式进行访问。用户还可以利用ENA提供的新服务（http: //www.ebi.ac.uk/ena/search）进行序列相似性搜索。当然用户还可以通过EMBL数据库提供的FTP站点（ftp: //ftp.edi.ac.uk/pub/databases/embl/）和SRA和测序数据库提供的FTP站点（ftp: //ftp.sra.ebi.ac.uk/）进行批量下载。

3. DDBJ数据库

日本DNA数据库DDBJ（DNA data bank of Japan，http: //www.ddbj.nig.ac.jp/）于1984年建立，由信息生物学中心和国家遗传研究所的日本DNA数据库（CIB-DDBJ）维护，是世界三大DNA 数据库之一，也是亚洲唯一的核酸序列数据库。它首先反映日本产生的DNA数据，同时每天将收集的数据与EMBL-Bank和GenBank数据库进行交换。DDBJ的主要目标是提高国际核酸序列数据库（international nucleotide sequence database，INSD）的质量。90%的日本研究者的数据是通过DDBJ提交的（图1-4）。

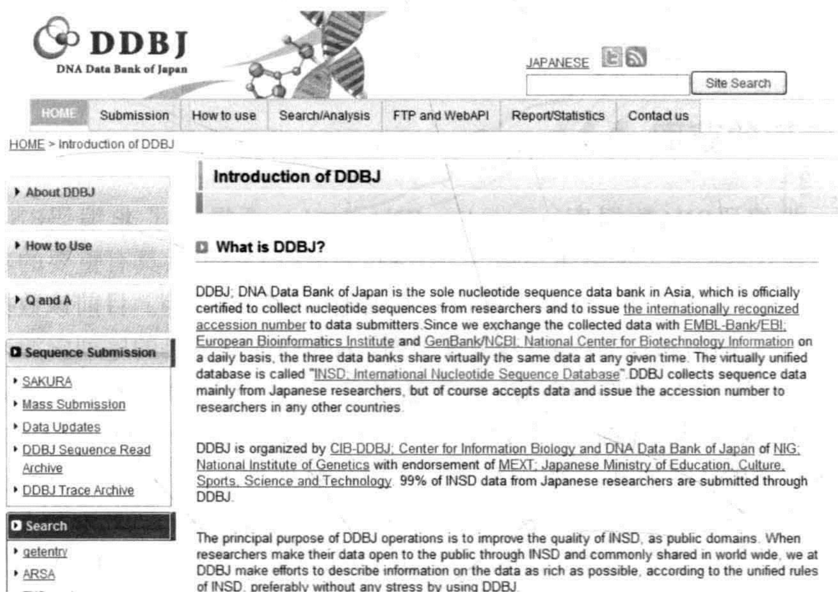


图1-4　DDBJ数据库的主页

（二）编码和非编码的DNA序列数据库

1. RefSeq　NCBI的参考序列（reference sequence，RefSeq）数据库（http: //www.ncbi.nlm.nih.gov/RefSeq/）为多种生物提供校正的序列数据信息及相关资料，用于医学、基因功能和基因功能比较研究。RefSeq数据库是一个综合的、非冗余的和具有较好注释的序列集合，包括基因组序列、RNA序列和蛋白质序列。RefSeq数据库区别于其他数据库的主要特征包括非冗余性、明确的核酸和蛋白质序列的对应关系、实时更新和数据的证实、格式的一致和截然不同的Accession号等。截止到2011年7月，在RefSeq数据库发布的49版中共存储了2.4百万个基因组记录、2.6百万个RNA记录和13.1百万个蛋白质记录。

2. PseudoGene　假基因数据库（PseudoGene，http: //www.pseudogene.org/）由耶鲁大学建立的一个存储真核生物和原核生物基因组中的假基因信息的综合数据库。PseudoGene数据库提供了友好的用户界面，根据不同的假基因特征将数据分为不同的部分，同时允许用户使用感兴趣的关键字进行查找和下载服务。

3. STRBase 短串联重复DNA数据库(short tandem repeat DNA internet database，STRBase)由位于美国马里兰州的国家标准与技术局的John M.Butler等人于1997年建立(http://www.cstl.nist.gov/div831/strbase/)。STRBase主要包括四个部分：一般数据、法医STR信息、其他的DNA标记信息和非人类DNA资源以及其他资源和工具。一般数据部分是STRBase的核心，包括一些广泛使用的STR标记的数据。这些数据以"STR资料概览"的形式被展示。这些资料概览由四个部分组成：一般信息、PCR引物、PCR产物大小和其他数据。一般信息区段包括STR位点的其他名字，它在染色体上的位置，核心STR重复单位的序列，它的GenBank序列号和参考序列中重复单位的个数。

4. TRDB 串联重复数据库(tandem repeats database，TRDB)由波士顿大学生物计算和信息中心的Gary Benson于2006年建立(http://tandem.bu.edu/cgi-bin/trdb/trdb.exe)。TRDB数据库收录了基因组DNA序列中的串联重复序列和各种分析工具。TRDB数据库提供了一系列服务包括：串联重复序列查找工具的下载，查询和过滤服务，基于序列相似性的重复序列聚类、多态的预测，PCR引物的选择和数据的下载。

二、RNA序列的获取

1. ncRNAdb 非编码RNA数据库(noncoding RNA database)提供了非编码RNA的序列和功能信息。虽然这些RNA不编码蛋白质，但是这些非编码RNA仍然具有重要的功能包括染色质结构重建、基因表达的转录和翻译调控和亚细胞位置的调控等。目前该数据库收录了来自99种真核生物、细菌和古细菌的3万多条序列。ncRNAdb的主要的序列资源来自于GenBank。还有一部分鼠和人类的ncRNA注释信息来自于FANTOM3数据库(http://fantom.gsc.riken.jp/4/)和H-Iev人类基因综合注释数据库(http://jbirc.jbic.or.jp/hinv/ahg-db/index.jsp)。细菌的小细胞质RNA序列和注释信息来自于Rfam数据库(http://rfam.sanger.ac.uk/)。ncRNAdb中的数据可以通过以下几种方法进行检索：Search(http://ncrnadb.trna.ibch.poznan.pl/search.html)，BLAST(http://ncrnadb.trna.ibch.poznan.pl/blast.html)，Browse(http://ncrnadb.trna.ibch.poznan.pl/Browser.html)，Download(http://ncrnadb.trna.ibch.poznan.pl/download.html)。

2. Rfam Rfam是通过多序列比对、二级结构和方差模型方法建立的非编码RNA家族数据库。Rfam可以通过位于英国的http://www.sanger.ac.uk/Software/Rfam/或者位于美国的http://rfam.wustl.edu/站点进行访问。Rfam数据库可以分为三个主要的功能类：非编码RNA基因、结构化的顺式调控元件和自主剪切的RNA。这些具有功能的RNA二级结构往往比RNA序列更保守。Rfam发布的第1版仅包含25个家族的5万个非编码RNA基因。截止到2011年6月Rfam发布了第十个版本包含1973个家族。

3. GtRDB 基因组tRNA数据库(genomic tRNA database，GtRDB)(http://gtrnadb.ucsc.edu)存储了已完成和接近完成的基因组中由tRNAscan-SE程序预测的tRNA基因。截止到2011年4月，GtRDB数据库包含了来自46种真核生物、86种古细菌和629种细菌的tRNA基因。

4. miRBase miRBase(http://www.mirbase.org/)是一个主要的存储所有在科学文献中发表的微小RNA(microRNA)序列和注释的国际数据库。miRBase建立于2002年，截止到2011年4月已经发布了第17版本，包含了140多个物种的1.6万个miRNA记录。图1-5显示了miRbase中miRNA记录和在Pubmed中关于miRNA文献数目的增长趋势。miRBase数据库主

要为用户提供以下几种服务：①可以检索已经公开发表的miRNA序列和注释信息；②可以获得和下载miRNA的发卡和成熟序列，也可以通过网页（http://www.mirbase.org/ftp.shtml）下载miRBase中的所有序列和注释信息；③miRBase Registry（http://www.mirbase.org/registry.shtml）允许用户提交新发现的miRNA，并提供专有的名称；④用户可以通过miRBase数据库连接到microCom获得预测的靶基因。

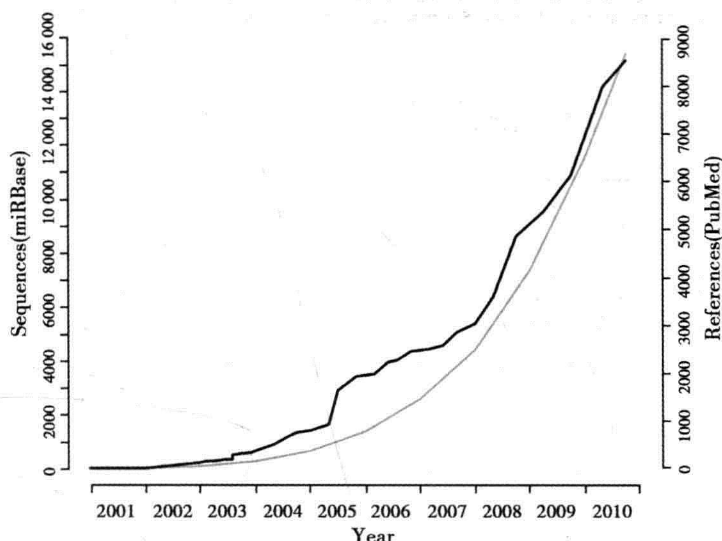


图1-5　miRBase中miRNA记录和在Pubmed中关于miRNA文献数目的增长趋势
来源于Kozomara A, Griffiths-Jones S.miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res,2001;39:152-157.

5. UTRdb/UTRsite　UTRdb是真核生物mRNA的5'端和3'端非翻译区序列的非冗余数据库，UTRsite搜集这些非翻译区序列中的功能片段（http://utrdb.ba.itb.cnr.it/）。UTRdb/UTRsite数据库现在主要分为两个部分UTRef 和UTRfull。UTRef部分收录了来自于79个物种48.3万个基因的47.3万条5' UTR和52.7万条3' UTR记录，同时还存储了78.8万个UTRsite模体、2万个实验验证的miRNA的靶点和24.2万个保守区域。UTRfull部分主要针对人类，包括了来自于ASPicDB（http://t.caspur.it/ASPicDB/index.php）数据库全长转录本的非翻译区序列，12.4万个5' UTR和19.4万个3' UTR，64.9万个保守元件和10.5万个实验验证的miRNA靶点。

三、蛋白质序列的获取

1. NCBI Protein database　NCBI的Entrez蛋白数据库（http://www.ncbi.nlm.nih.gov/sites/entrez?db=protein）整合了来自于多种资源的蛋白质序列，这些资源包括SwissProt，the Protein Information Resource，the Protein Research Foundation，the Protein Data Bank和从GenBank和RefSeq数据库中有注释的编码区直接翻译得到的蛋白质序列。通过Entrez蛋白数据库中的蛋白质序列记录还可以查看相关的预处理的蛋白序列BLAST比对结果，蛋白质结构，保守的蛋白结构域，核酸序列，基因组和基因。例如，要检索人类发状分裂相关增强子-5（hairy and enhancer of

split 5）的蛋白质序列，可以在输入栏中输入检索条件"(hairy and enhancer of split 5[Protein Name]) AND human[Organism]"，然后点击"Search"检索，检索的页面如图1-6所示。

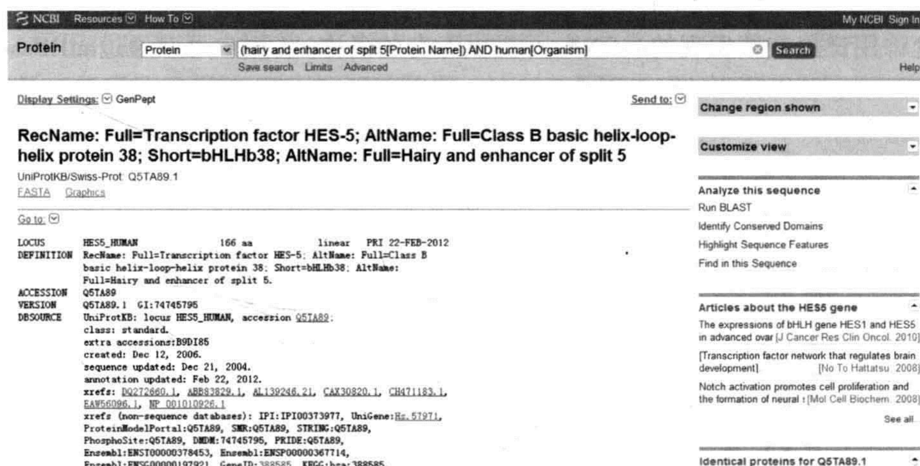


图1-6 在Entrez中检索人发状分裂相关增强子-5蛋白数据

2. EXProt EXProt(database for experimentally verified protein functions, http: //www.cmbi.kun.nl/EXProt/)是一个非冗余的蛋白质数据库，只存储那些在基因组注释计划和其他公共数据库中功能得到实验证实的那些蛋白质。EXProt发布的2.01版本中包括了6491条记录。这些记录来自于大肠埃希菌基因组的PseudoCAP(http: //www.pseudomonas.com/)计划和蛋白质组数据库GenProtEC(http: //genprotec.mbl.edu/)还有EMBL核酸序列数据库的原核生物部分。在EXProt中的记录都有一个唯一的ID号和相对应的来源物种，蛋白序列、功能注释，来源数据库、对应的基因名字和在PubMed相关的文献。

3. MIPS 数据库(http: //mips.gsf.de/)由德国慕尼黑蛋白质序列信息中心(databases at Munich information center for protein sequences, http: //www.helmholtz-muenchen.de/en/ibis)建立和维护。MIPS支持和维护一系列的基因组数据库以及提供微生物、真菌和植物基因组的系统的比较基因组学分析服务。同时该站点还提供基因组分析工具、数据库检索服务、表达分析、蛋白质互作等网络服务。

4. PIR 蛋白质信息数据库(protein information resource, PIR) (http: //pir.georgetown.edu/)是由美国国家生物医学研究基金会NBRF(national biomedical research foundation)于1984年建立的一个综合公共生物信息资源，其目的是支持基因组、蛋白质组和系统生物学的研究，帮助研究者鉴别和解释蛋白质序列信息，研究分子进化、功能基因组，进行生物信息学分析。PIR数据库除了提供蛋白质的序列数据外，还包括以下的信息：蛋白质名称、分类、来源、原始数据的参考文献、蛋白质功能和蛋白质一般特征，序列中相关位点和功能区域。PIR还提供了超家族、域和模体水平上的蛋白分类。此外PIR还提供了三种类型的检索服务包括基于文本的交互式检索、序列相似性检索和综合序列相似性，注释信息和蛋白质家族信息的高级检索。在PIR的站点上也提供了常规的生物信息学工具，进行更深入的数据发掘。PIR现在已经与Swiss-Prot和 TrEMBL合作，共同构成了UniProt数据库。

5. Swiss-Prot Swiss-Prot(UniProt/Swiss-Prot) (http: //www.expasy.org/sprot)由Geneva大学和欧洲生物信息学研究所(EBI)于1986年联合建立的，它是目前国际上权威的蛋白质序列

数据库。数据库由蛋白质序列条目构成，每个条目包含蛋白质序列、引用文献信息、分类学信息、注释等，注释中包括蛋白质的功能、转录后修饰、特殊位点和区域、二级结构、四级结构、与其他序列的相似性、序列残缺与疾病的关系、序列变异体和冲突等信息。SWISS-PROT中尽可能减少了冗余序列，并与其他30多个数据库建立了交叉引用，其中包括核酸序列库、蛋白质序列库和蛋白质结构库等。Swiss-Prot中的数据主要来源于：①从核酸数据库经过翻译推导而来；②从蛋白质数据库PIR挑选出合适的数据；③从科学文献中摘录；④研究人员直接提交的蛋白质序列数据。Swiss-Prot 在 2011年9月发布的第2011_09版本中存储了来自于20.1万篇参考文献的53.2万条序列记录包括了1.8亿个氨基酸（图1-7）。

Swiss-Prot数据库与其他蛋白质数据库相比较具有三个明显的特点：①在Swiss-Prot数据库中每一个序列记录包括核心数据和注释两大类。核心数据包括序列、参考文献和分类信息等。而注释包括功能描述、翻译后修饰、结构域和功能位点、蛋白质的四级结构、与该蛋白质相关的疾病和序列的变化信息等。②Swiss-Prot数据库尽量将相关的数据合并，降低数据的冗余度。如果不同来源的原始数据有矛盾，则在相应的序列特征表中加以注释。③Swiss-Prot目前已经建立了与其他30多个相关数据库的交叉索引，便于用户迅速得到在其他数据库中的相关信息。

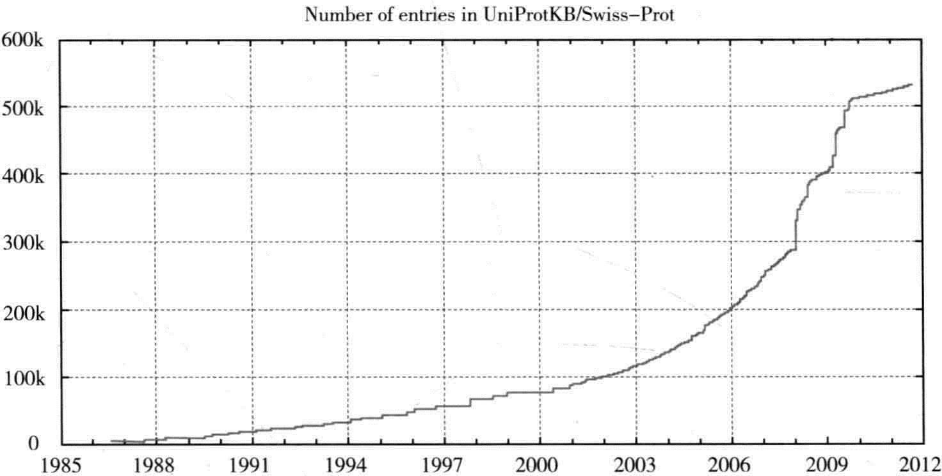


图1-7 Swiss-Prot数据库中记录的数目和增长趋势

来源于：http://web.expasy.org/docs/relnotes/relstat.html

第二节

双序列比对

Section 2 Pairwise Sequence Alignment

一、序列比对的相关概念

(一) 相似性与同源性

序列比对(sequence alignment)是通过在序列中搜索一系列单个性状或性状模式来比较2个(双序列比对)或更多(多重序列比对)序列的方法。序列比对最根本目的之一是通过对比不同物种序列的相似性判断它们之间是否具有同源性。值得注意的是,相似性(similarity)和同源性(homology)虽然在某种程度上具有一致性,但它们是完全不同的两个概念。相似性和同源性是序列比较和分析的基础。同源序列,简单地说,是指从某一共同祖先经趋异进化而形成的不同序列。相似性是指序列比对过程中用来描述检测序列和目标序列之间相同DNA碱基或氨基酸残基顺序所占比例的高低。同源性是序列同源或者不同源的一种论断,是个定性的概念,没有度的差异,而相似性是两个序列相关性的量化。两条序列之间要么是同源的,要么是不同源的,决不能像相似性那样具有多或少的数量关系,例如,不能说两条序列之间有90%的同源。

如果两个DNA序列经过序列比对具有较高的相似性,则检测序列和目标序列可能是同源序列;而当相似性程度低于20%时,就难以确定或者根本无法确定其是否具有同源性。同源序列可进一步分为两种:直系同源(orthology)和旁系同源(paralogy)。直系同源是指在种系形成(speciation)过程中起源于一个共同祖先的不同种系中的DNA或蛋白质序列。若一个基因原先存在于某个物种,而该物种分化为了两个物种,那么新物种中的基因是直系同源的;旁系同源的序列因基因复制(gene duplication)而被区分开。若生物体中的某个基因被复制了,那么两个副本序列就是旁系同源的。直系同源的一对序列称为直系同源体(orthologs),旁系同源的一对序列称为旁系同源体(paralogs)。直系同源体通常有相同或相似的功能,但对旁系同源体则不一定:由于缺乏原始的自然选择的力量,繁殖出的基因副本可以自由的变异并获得新的功能。

(二) 空位罚分概念及策略

一般在进行双序列或者多序列比对时为了获得两个或多个序列的最佳比对要对序列插入空位(gap)。空位是指在进行序列比对时,为了获得最佳比对结果,算法权衡后在两条或

多条序列中产生的间隔区。引入空位的数量和位置对比对结果有显著影响，因此必须在比对计分时对其罚分。空位罚分（gap penalty）指序列比对分析时为了反映核酸或氨基酸的插入或缺失等情况而插入空位并进行罚分，以控制空位插入的合理性。

除了对应于单字符插入和删除的空位，比对中还经常用到更大的对应于多个连续字符插入和删除的空位。多个连续字符的插入和删除可由多次独立的单字符插入和删除造成，也可由一次多字符插入和删除造成。尽管单字符突变的发生率高于多字符突变的发生率，从概率上说，引起一次多字符插入和删除的概率要大于引起多次独立单字符插入和删除的概率。此外，对于长的空位，它们出现在序列的头、中和尾也常常具有不同意义。最优的序列比对通常具有以下两个特征：尽可能多的匹配和尽可能少的空位。插入任意多的空位可能会产生较高的分数，但找到的并不一定是真正相似的序列。

有2个参数应用于空位罚分设定，一个与空位设置（gap opening）有关，另一个与空位扩展（gap extension）有关（表1-5）。任一空位的出现均处以空位设置罚分，而任一空位的扩大必须处以空位扩展罚分。对于一个空位长度为k的罚分W_K可用下式表示：

$$w_K = a + bk \tag{1-1}$$

其中a是空位设置罚分，b为空位扩展罚分。这两个参数值设置的变化对联配产生影响。

表1-5　空位设置和空位扩展罚分对联配的影响

空位设置罚分	空位扩展罚分	说明
大	大	极少插入或缺失：适用于非常相关蛋白质间的联配
大	小	少量大块插入：用于整个功能与可能插入的情况
小	大	大量小块插入：适用于亲缘关系较远的蛋白质同源性分析

（三）替换记分矩阵

对于序列中的插入和删除突变，序列比对采用插入空位来处理，使得原本对应的字符仍旧能够对应；而对于序列中的替换突变，需要考虑不同替换的意义。合理而精确的记分需要考虑替换的各种情形。对于DNA和RNA序列，情况特别简单，施用于4种碱基和6种彼此间替换关系的记分规则可用简单的替换记分矩阵来描述。对于蛋白质序列，因为蛋白质由20种氨基酸构成，且不同的氨基酸具有不同的理化性质，情况较为复杂，存在许多不同的替换记分矩阵。

由于替换有多种情形，且可按不同方式罚分，如何精确处理序列中的替换突变十分重要。显然，不同字符间的替换具有不同的概率，也具有不同的意义；同时，不同物种间的替换也有不同的概率和意义。精确地处理替换需要考虑各种情形，而方便地处理替换则要求把不同的处理方法参数化，这些参数就是替换记分矩阵，它们定量地标示了不同替换的意义。

1. DNA序列比对的替换记分矩阵

（1）等价矩阵：等价矩阵（表1-6）是最简单的一种替换记分矩阵，其中，相同核苷酸间的匹配得分为1，不同核苷酸间的替换得分为0。尽管含义清晰明了，由于不含有碱基的任何理化信息和不区别对待不同的替换，在实际的序列比对中较少使用。

表1-6 DNA等价矩阵

	A	T	C	G
A	1	0	0	0
T	0	1	0	0
C	0	0	1	0
G	0	0	0	1

（2）BLAST矩阵：经过大量实际比对发现，如果令被比对的两个核苷酸相同时得分为+5，反之得分为–4，则比对效果较好。表1–7是其替换记分矩阵，这个矩阵广泛地被DNA序列比对所采用，称为BLAST矩阵。BLAST是目前最流行的核酸序列数据库搜索程序。

表1-7 BLAST矩阵

	A	T	C	G
A	5	–4	–4	–4
T	–4	5	–4	–4
C	–4	–4	5	–4
G	–4	–4	–4	5

（3）转换–颠换矩阵（transition–transversion matrix）：核酸的碱基按照环结构特征被划分为两类，一类是嘌呤（腺嘌呤A、鸟嘌呤G），它们有两个环；另一类是嘧啶（胞嘧啶C、胸腺嘧啶T），它们只有一个环。如果DNA碱基的替换保持环数不变，则称为转换，如A→G、C→T；如果环数发生变化，则称为颠换，如A→C、A→T 等。在进化过程中，转换发生的频率远比颠换高，表1–8 所示的矩阵用来反映这种情况，其中转换的得分为–1，而颠换的得分为–5。

表1-8 转换–颠换矩阵

	A	T	C	G
A	1	–5	–5	–1
T	–5	1	–1	–5
C	–5	–1	1	–5
G	–1	–5	–5	1

2. 蛋白质序列比对的替换记分矩阵　对于蛋白质序列，记分矩阵主要用于记录在做序列比对时两个相对应的残基的相似度。简单的替换记分办法，如+1表示匹配，0表示失配，是不够的。构成蛋白质的氨基酸具有不同的生物化学特性，这些特性可影响它们在进化过程中的相互替换。下面介绍两种常用的氨基酸替换记分矩阵。

（1）PAM矩阵：对于氨基酸之间的替换，对实际替换率的直接观察常常是导出合理的记分的好方法，由此产生的一组替换记分矩阵是点突变可接受矩阵（point accepted matrix，PAM）。它们基于氨基酸进化的点突变模型，即如果两种氨基酸替换频繁，说明自然界易接

受这种替换，那么这对氨基酸替换得分就应该高。PAM矩阵是目前蛋白质序列比对中最广泛使用的记分方法之一，1个PAM的进化距离表示在100个残基中发生一个可以接受的残基突变的概率。对应于一个更大进化距离间隔的突变矩阵，可以通过对原始矩阵进行一定的数学处理获得。将PAM-1自乘*n*次，可以得到PAM-*n*。例如，PAM250相似性分数矩阵（表1-9）相当于在两个序列之间具有20%的残基匹配。对于PAM-*n*矩阵，*n*越小表示氨基酸变异的可能性越小，高相似序列之间的比对应该选用*n*值小的矩阵，低相似序列之间的比对应该选用*n*值大的矩阵。

表1-9　PAM-250矩阵

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0	0	0
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2	-1	0
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2	2	1
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2	3	3
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2	-4	-5
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2	1	3
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2	3	3
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1	0	0
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2	1	2
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4	-2	-2
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2	-3	-3
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2	1	0
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2	-2	-2
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1	-4	-5
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1	-1	0
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1	0	0
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0	0	-1
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6	-5	-6
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2	-3	-4
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	-2	-2
B	0	-1	2	3	-4	1	3	0	1	-2	-3	1	-2	-4	-1	0	0	-5	-3	-2	3	2
Z	0	0	1	3	-5	3	3	0	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2	2	3

PAM矩阵的制作步骤是：

1）构建序列相似（大于85%）的比对。

2）计算氨基酸*j*的相对突变率*mj*（j被其他氨基酸替换的次数）。

3）针对每个氨基酸对*i*和*j*，计算*j*被*i*替换的次数。

4）替换次数除以相对突变率（*mj*）。

5）利用每个氨基酸出现的频度对*j*进行标准化。

6）取常用对数，得到PAM-*i*（*i*, *j*）。

（2）BLOSUM矩阵：BLOSUM（block substitution matrix）矩阵由Henikoff夫妇从蛋白质模块数据库BLOCKS中找出的另一种氨基酸替换记分矩阵，用于解决序列的远距离相关。在构建矩阵过程中，通过设置最小相同残基数百分比将序列片段整合在一起，以避免由于同一个残基对被重复计数而引起的任何潜在偏差。在每一片段中，计算出每个残基位置的平均贡献，使得整个片段可以有效地被看做为单一序列，通过设置不同的百分比，产生了不同矩阵。表1-10所示的BLOSUM矩阵是由具有62%相同比例的序列被组合统计后形成的矩阵。注意，在比对高度相似的序列时使用较高值的矩阵（高至BLOSUM-90），在比对差异大的序列时使用较低值的矩阵（低至BLOSUM-30）。对于BLOSUM-*n*矩阵，*n*越小则表示氨基酸相似的可能性越小，高相似的序列之间比较应该选用*n*值大的矩阵，低相似序列之间的比对应该选用*n*值小的矩阵。例如，BLOSUM-62用来比较62%相似度的序列，BLOSUM-80用来比较80%左右的序列。

表1-10 BLOSUM-62矩阵

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4

二、双序列比对算法

生物序列(DNA序列、RNA序列和蛋白质序列)可以看作是由固定的字母表中的字母所组成的字符串,两条序列s和t的比对可以简单的表示为:把s和t这两条序列上下排列起来,在某些位置插入空位,然后依次比较它们在每一个位置上字符的匹配情况,从而找出使这两条序列产生最大相似度得分的排列方式和空位插入方式。

(一)点阵图法

点阵(dot matrix)分析是一种简单的图形显示序列相似性的方法。将两条待比较的序列分别放在矩阵的X/Y轴上,从下往上和从左到右比较,当对应行与列的字符匹配时,则在矩阵对应的位置上打点。逐个比较所有的字符对,最终形成一个点矩阵。点阵图可以应用于自身比对,用来寻找序列中的正向或反向重复序列,查找蛋白质的重复结构域,相同残基重复出现的低复杂区和RNA二级结构中的互补区域。同时点阵图也可以对两条序列的相似性做整体的估计。点阵分析具有直观性和整体性的优点,而且不依赖于空位参数,可以寻找两序列间所有可能的残基匹配。点阵分析允许随时动态地改变最高和最低界限值,可以用来搜索区分信号和背景标准的严格程度。总之,点阵分析不依赖任何先决条件,是一种可用于初步分析的理想工具。但是点阵分析具有不能很好地兼容打分矩阵、滑动窗口和与阈值的选择过于经验化、信噪比低和不适合进行高通量数据分析等缺点。常用的点阵分析工具见表1-11。

表1-11 常用的点阵分析工具

工具名	网址	备注
DNA Strider	http://www.cellbiol.com/soft.htm	Mac
Dotter	http://sonnhammer.sbc.su.se/Dotter.html	Unix/Linux, X-Windows
Dotlet	http://myhits.isb-sib.ch/cgi-bin/dotlet	Web
DNAdot	http://arbl.cvmbs.colostate.edu/molkit/dnadot/	Web

(二)动态规划算法

对于两条序列的比对问题人们提出了很多算法,其中基于动态规划的算法是目前最基本的算法。1970年Saul Needleman 和 Christian Wunsch两人首先将动态规划算法用于两条序列的全局比对。全局比对是指将参与比对的两条序列里面的所有字符进行比对。全局比对主要被用来寻找关系密切的序列。后来, Temple Smith和Michael Waterman两人于1981年对双序列的局部比对进行了研究,产生了Smith-Waterman算法。这两种算法均可以用于核酸和蛋白质序列。在给定空位罚分和替换矩阵情况下,它们总是能给出具有最高(优)联配值的联配。但是,这个联配并不需要达到生物学意义上的显著水平。动态规划首先对于如下假定的序列:

(1)a, b是使用某一字符集Σ的序列(DNA或蛋白质序列);

（2）m = a的长度；

（3）n = b的长度；

（4）S（i, j）是按照某替换记分矩阵得到的前缀a[$1...i$]与b[$1...j$]最大相似性得分；

（5）w（c，d）是字符c和d按照替换记分矩阵计算的得分。

可按照某种记分规则建立得分矩阵：

$S(i,0)=0,\quad 0 \leqslant i \leqslant m$

$S(0,j)=0,\quad 0 \leqslant j \leqslant n$

$$S(i,j)=\max\begin{cases}0 & \\ S(i-1,j-1)+w(a_i,b_j) & \text{匹配或错配} \\ S(i-1,j)+w(a_i,-) & \text{插入} \\ S(i,j-1)+w(-,b_j) & \text{缺失}\end{cases} \qquad (1\text{-}2)$$

例如，对于序列a=ACACACTA，序列b=AGCACACA，记分规则w（匹配）=+2；w（a，–）=w（–，b）=w（失配）=–1，则获得的得分矩阵如图1–8所示。接着，反向搜寻最大得分，同时记下读取路径。为了得到最佳比对，必须从得分最高的位置S（i, j）开始，在矩阵的（i–1，j），（i，j–1）或（i–1，j–1）位置中寻找下一个最大得分位置，记下路径（画箭头），当两个（或三个）位置得分相等时，取对角线方向，依此规则搜寻，直至到起点（0，0）。在本例中，最大得分对应的位置分别为（8，8）、（7，7）、（7，6）、（6，5）、（5，4）、（4，3）、（3，2）、（2，1）、（1，1）和（0，0）（图1–9）。

S=

	–	A	C	A	C	A	C	T	A
–	0	0	0	0	0	0	0	0	0
A	0	2	1	2	1	2	1	0	2
G	0	1	1	1	1	1	1	0	1
C	0	0	3	2	3	2	3	2	1
A	0	2	2	5	4	5	4	3	4
C	0	1	4	4	7	6	7	6	5
A	0	2	3	6	6	9	8	7	8
C	0	1	4	5	8	8	11	10	9
A	0	2	3	6	7	10	10	10	12

图1–8 一个得分矩阵实例

S=

	–	A	C	A	C	A	C	T	A
–	0	0	0	0	0	0	0	0	0
A	0	2	1	2	1	2	1	0	2
G	0	1	1	1	1	1	1	0	1
C	0	0	3	2	3	2	3	2	1
A	0	2	2	5	4	5	4	3	4
C	0	1	4	4	7	6	7	6	5
A	0	2	3	6	6	9	8	7	8
C	0	1	4	5	8	8	11	10	9
A	0	2	3	6	7	10	10	10	12

图1–9 得分矩阵路径实例

最后构建最佳匹配。在读取路径中要求：对角线对应匹配（或失配）、上下箭头对应删除、左右箭头对应插入。依此规则，我们可以得到本例的最佳匹配为：

序列a =	A	–	C	A	C	A	C	T	A
序列b =	A	G	C	A	C	A	C	–	A

现在看算法的复杂度。从所使用的数据结构本身及其计算过程来看，序列两两比对基本算法的空间复杂度和时间复杂度都是0（mn）。

动态规划算法大致包括：①按照规则建立得分矩阵；②反向读取最大得分，构建最佳匹配。每一步都包括若干子步骤。按照规则建立得分矩阵的流程是：

```
for i=0 to length( A )
    F( i,0 )←0
  for j=0 to length( B )
    F( 0,j )←0
for i=1 to length( A )
   for j= 1 to length( B )
   {
     Choice1 ← F( i–1,j–1 )+ S( A( i ), B( j ))
     Choice2 ← F( i–1,j )+ d
     Choice3 ← F( i,j–1 )+ d
     F( i,j )← max( Choice1, Choice2, Choice3 )
   }
```

反向读取最大得分,构建最佳匹配流程是:

```
AlignmentA ← ""
  AlignmentB ← ""
  i ← length( A )
  j ← length( B )
  while ( i > 0 and j > 0 )
{
    Score ← F( i,j )
    ScoreDiag ← F( i – 1,j – 1 )
    ScoreUp ← F( i,j – 1 )
    ScoreLeft ← F( i – 1,j )
    if ( Score == ScoreDiag + S( A( i–1 ), B( j–1 )))
{
      AlignmentA ← A( i–1 )+ AlignmentA
      AlignmentB ← B( j–1 )+ AlignmentB
      i ← i – 1
      j ← j – 1
}
  else if ( Score == ScoreLeft + d )
{
      AlignmentA ← A( i–1 )+ AlignmentA
      AlignmentB ← "–" + AlignmentB
      i ← i – 1
}
    otherwise ( Score == ScoreUp + d )
```

```
{
    AlignmentA ← "-" + AlignmentA
    AlignmentB ← B(j-1) + AlignmentB
    j ← j - 1
}
}
```

(三) 基于双序列比对的数据库搜索

一条序列与整个数据库中所有序列比对的过程可以看做是双序列比对的扩展。本质上这与两条序列的比较没有什么两样，只是要重复成千上万次。但是要快速实现数据库的搜索并非易事。无论Needleman-Wunsch的算法还是Smith-Waterman算法，对于数量不大的序列来说其运行时间上可接受。对于大规模的数据库搜索，它们都非常耗时，所以必须考虑在一个合理时间内完成搜索比较操作。FASTA和BLAST是目前基于局部相似性的数据库搜索程序。BLAST(basic local alignment search tool，基本局部联配搜索工具)是基于匹配短序列片段，用一种强有力的统计模型来确定未知序列与数据库序列的最佳局部联配。BLAST算法本身很简单，它的基本要点是序列片段对(segment pair)的概念。所谓序列片段对是指两个给定序列中的一对子序列，它们的长度相等，并且可以形成无空位的完全匹配。BLAST首先找出探测序列和目标序列间所有匹配程度(以得分计)超过一定阈值的序列片段对，然后对片段对根据给定的相似性阈值进行延伸，得到一定长度的相似性片段，最后给出高分值片段对(high-scoring pairs，HSPs)。改进后的BLAST允许空位的插入。BLAST实际上是综合在一起的一组程序，不仅可用于直接对蛋白质序列数据库和核酸序列数据库进行搜索，而且可以将探测序列翻译成蛋白质后再进行搜索，以提高搜索结果的灵敏度。

大多数研究目前都通过国际互联网Internet应用NCBI研制的BLAST程序来进行DNA和蛋白质序列相似性搜索。用一组BLAST程序联配可以快速进行核酸和蛋白质序列库的相似性检索。采用BLAST的基本算法编成了若干个不同的程序，分别使用特定的序列库和用于特定类型的输入序列。BLAST家族包含的成员很多，提供各种不同需要的比对分析，最常见也是最重要的五个成员分别是blastn、blastp、blastx、tblastn和tblastx(表1-12)。下面以刚才的人类发状分裂相关增强子-5蛋白为例，说明如何通过NCBI网页搜索该蛋白的同源序列，步骤如下:

1. 下载人类发状分裂相关增强子-5蛋白序列。
2. 登录NCBI主页http: //www.ncbi.nlm.nih.gov/。
3. 点击“BLAST”。
4. 对话框中输入人类发状分裂相关增强子-5蛋白序列。
5. 选择蛋白质数据库Non-redundant protein sequences(nr) (图1-10)。
6. 其他参数使用默认参数。
7. 点击BLAST按钮，得到数据库搜索的结果(图1-11)。点击感兴趣的序列可以得到序列匹配的详细界面。

表1-12　BLAST家族常用的工具

程序名	查询序列	数据库	搜索方法
Blastn	核酸	核酸	用查询序列逐一搜索数据库中的序列
Blastp	蛋白质	蛋白质	用查询序列逐一搜索蛋白质数据库中的序列
Blastx	核酸	蛋白质	将核酸序列翻译成蛋白质序列后逐一搜索蛋白质数据库中的序列
TBlastn	蛋白质	核酸	将查询蛋白质序列逐一搜索核算数据库中的核酸序列翻译后的蛋白质序列
TBlastx	核酸	核酸	将核酸序列翻译成蛋白质序列后逐一搜索核酸数据库中的核酸序列翻译后的蛋白质序列

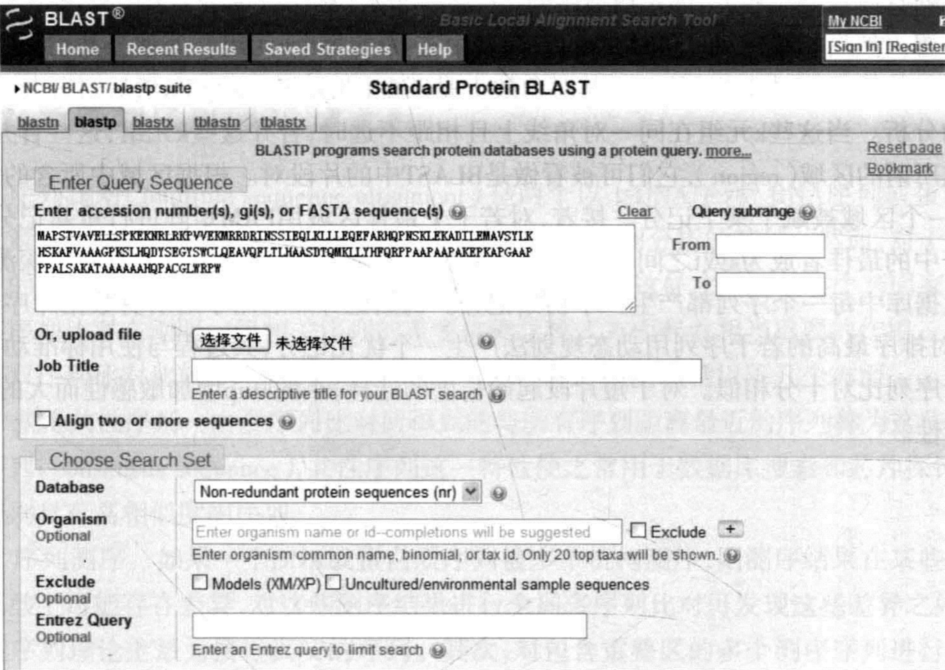


图1-10　输入BLAST查询序列、选择数据库

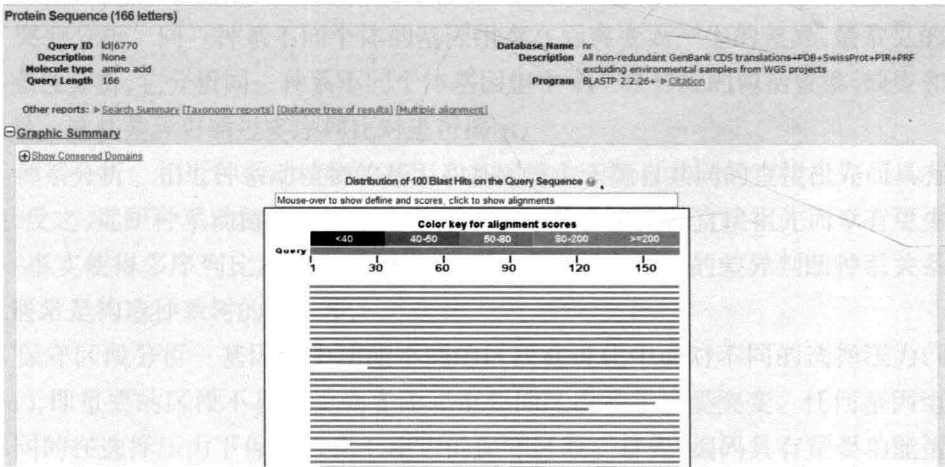


图1-11　BLAST查询返回结果图示

网络版本的Blast服务具有方便，容易操作，数据库同步更新等优点，缺点是不利于操作大批量的数据，同时也不能自己定义搜索的数据库，只能对NCBI所提供的数据库进行序列相似性分析；NCBI同时提供了可本地化安装的Blast软件包，此本地化的软件包允许用户在自己的计算机上安装Blast系统，并可构建自己的数据库，大大提高了同源性分析的准确性和一致性。

FASTA是另一个常用的基于局部比对的数据库搜索程序，算法是由Lipman和Pearson于1985年发表的。假定探测序列s和目标序列t是两个被比对的蛋白质序列，且长度分别为|s|=m，|t|=n。FASTA进行的比较由确定两者公共的（即相匹配的）短片段开始，这些短片段称为k元组（k-tuple），短片段的起始长度k一般是1~2个氨基酸（k的值是一个参数，称ktup，对于DNA序列它通常要大些）。首先扫描序列s，产生一个表（称作查询表），表中列出s中k元组的所有位置。然后扫描序列t，同时在s的查询表中查找t的每个k元组。处理公共出现的k元组的结果是构造一个动态规划矩阵，其对角线上是匹配的k元组。FASTA然后对公共k元组作详细的分析。当这些k元组在同一对角线上且相距不远时，合并这些k元组，这些合并的k元组构成所谓的区域（region），它们可被看做是BLAST中的片段对。根据区域中所含的匹配或失配，一个区域被赋予某个记分。接着，对若干个最佳区域用PAM矩阵进行重新记分，这些新记分中的最佳者成为s或t之间相似性的一个初步度量，称作初始记分。对一个探测序列，它与数据库中每一个序列都产生一个初始记分。初始记分然后用于对所有数据库序列进行排序，对排序最高的若干序列用动态规划法产生一个优化记分，其过程与使用标准动态规划法进行序列比对十分相似。对于短片段起始长度的选择，小的ktup增加敏感性而大的ktup增加特异性。

第三节

多序列比对

Section 3 Multiple Sequence Alignment

一、多序列比对简介

多序列比对(multiple sequence alignment)是两个以上DNA序列、RNA序列或蛋白质序列的比对,目标是发现多条序列的共性。双序列比对是序列分析的基础。然而,对于构成基因家族的成组序列来说,我们要建立多个序列之间的关系,这样才能揭示整个基因家族的特征。多序列比对在阐明一组相关序列的重要生物学模式方面起着相当重要的作用。

与双序列比对比较,多序列比对具有更广泛的重要应用,包括以下几个方面:

1. 获得共性序列　由多序列比对所得到的与所有序列距离最近的序列称为这些序列的共性序列(consensus sequence),共性序列这一特性使之常用于数据库搜索和芯片探针设计,用于识别具有高相似度的序列。

2. 序列测序　如果一个DNA或蛋白质序列被多个机构测序,则测序结果在某些核苷酸或氨基酸上可能存在差异,对这些测序结果进行全局多序列比对可发现这些差异之处,形成的共性序列理论上最为接近真实的序列。其次,对包含重叠区的多个测序序列进行局部多序列比对可发现这些重叠区,实现测序序列的拼接。另外,一个类似的应用是由表达序列标签(expressed sequence tag, EST)组装较长的重叠群(contig)甚至完整的mRNA。

3. 突变分析　同一种系不同个体的基因组存在因突变而产生的差异,最常见的是单核苷酸多态性分析,它分析同一种系不同个体基因组中单个核苷酸的包括置换、缺失和插入在内的变异。这些差异可通过多序列比对进行揭示。

4. 种系分析　相近种系动植物的基因和基因组由于源自共同的直接祖先而具有高度的相似性,反之,远距种系动植物的基因和基因组由于源自不同的直接祖先而享有更少的相似性,这一事实使得多序列比对常常用于根据基因或基因组序列的差异判断种系关系。多序列比对通常是构造种系树的第一步。

5. 保守区段分析　基因组中功能不同的区段在进化中面对不同的选择压力(selective pressure),即重要的区段不易接受突变而非重要的区段易于接受突变。任何基因组都包含大量不同的在选择压力下保持进化上稳定的保守区段。首先,编码具有重要功能蛋白质的基因高度保守,基因中的外显子尤其保守。其次,大量的基因调节单元,例如启动子和增强

子，在不同种系中通常高度保守。此外，近年来发现许多非编码RNA也是非常保守的。多序列比对是找出进化上保守的这些区段的基本方法。

6. 基因和蛋白质功能分析　分子生物学和发育生物学实验是揭示基因和蛋白质功能的经典方法。在大量基因和蛋白质的功能得以揭示和更多基因和蛋白质的序列得以测定后，根据与功能已知的同源基因和蛋白质进行多序列比对来推断新基因和蛋白质的功能已成为越来越普遍的一个研究手段。

7. RNA和蛋白质结构分析　类似地，可使用多序列比对考察种系相近的RNA和蛋白质家族，通过结构已知的RNA和蛋白质推断未知RNA和蛋白质的结构。需要注意的是，核苷酸序列和氨基酸序列的进化速度比RNA结构和蛋白质结构的进化速度要快，因此仅凭多序列比对仍难以确定RNA和蛋白质的结构。例如，人β球蛋白（beta-globin）和肌球蛋白（myoglobin）只有25%的氨基酸序列相同，但两者的三维结构却几乎相同。

8. 基因组结构分析　多序列比对可用于整个基因组，揭示基因组的结构特征和进化特征。随着越来越多基因组的测序，多序列比对已频繁用于基因组结构分析中，最典型的应用是UCSC基因组浏览器和Ensembl基因组浏览器。

二、多序列比对的方法

这些年来，在生物信息学领域提出了许多关于多序列比对的算法，如动态规划算法、渐进策略算法、迭代法、基于一致性的方法、遗传算法、模拟退火算法、隐马尔可夫模型、星形比对和树形比对等多序列比对算法。

（一）动态规划算法

动态规划算法由Needle-man和Wunsch于1970年提出，最初用于求两个序列的最佳比对。当把动态规划的基本思想推广到多序列比对时就是所谓的N维动态规划算法。由于动态规划法的时间与空间复杂性太高，人们发展了该算法的多种变体使得它们能够在合理的时间内找到优化比对。变体之一是Altschul等在1989年引入的一个算法，它能极大地缩小k维动态规划表的搜索空间，其中心思想如下。首先，对k个序列的$\binom{k}{2}$个配对按动态规划法进行配对比对，由于一个k序列比对对应于k维空间动态规划表中的一个路径，这些配对比对可看作是k维空间中的这个路径在不同的2维空间中的投影。其次，在相应的2维空间中，可以限制投影所可能历经的空间，从而限制在原始的k维空间中寻找优化多序列比对历经的路径。第三，每个投影定义了原始k维空间的一个子空间，这些子空间的交汇包含了k个序列的优化比对。该算法通常采用SP函数计分，并使用动态规划法搜索子空间的交汇来找到多序列比对在k维空间中的路径。一个关键点是，需要确定一个将多序列比对投影成配对比对的开支上限，该开支上限的选择应能保证动态规划法找到k个序列的最优比对。在使用启发式方法确定配对比对的开支上限时，若比对的质量表明开支上限不够大，则应增大开支上限。但是，一味地增大开支上限并不能持续提高多序列比对的质量。由于该动态规划法的变体

本质上属于启发式算法，它有一切启发式算法所固有的缺陷，即当开支上限定得过小时它可能找不到最优比对，而当开支上限定得过大时它所耗费的时间可能与标准动态规划法所差无几。

(二) 渐进策略算法

渐进策略算法最早由Feng和Doolittle于1987年提出。渐进多序列比对首先使用动态规划法构造全部k个序列的 $\binom{k}{2}$ 个配对比对(pairwise alignment)，然后以计分最高的配对比对作为多序列比对的种子，按计分高低依次选择序列，逐渐向已构造的多序列比对中加入序列，形成一个树状结构的多序列比对结果。

渐进多序列比对需要三个步骤：第一，使用动态规划法构造每个序列的配对比对，包括ClusterW在内的许多比对算法在这一步使用距离矩阵而不是相似性矩阵来描述序列间的关联性；第二，由距离矩阵构造一棵指导树(guide tree)，树的两个主要特征是拓扑结构和分枝长度，它一般并不被当作是种系树，只反映了参与比对的多个序列如何相关联，用来确定向正在进行的多序列比对加入新序列的次序；第三，以计分最高的配对比对作为多序列比对的种子，根据指导树逐渐向多序列比对中加入序列。这种方法在质量尤其是计算速度上、存储空间及可比对的序列数目方面比动态规划算法更优良。在比对过程中遵循“一旦引入一个空位则始终保持这个空位”的原则。为了最大程度地残基匹配，比对过程中采用可接受的点突变矩阵PAM。不仅允许相同残基的匹配，而且允许相似残基的匹配。其缺点是不能保证比对的结果是数学上的最优化比对。首先，渐进多序列比对可能会被一些伪强的、实际上是坏的种子所误导。如果一开始选择的两条序列的配对比对与实际上的最优多序列比对不一致，那么初始的配对比对中的错误在整个多序列比对构造过程中将始终存在并持续传播。其次，在比对的任何阶段出现失配时(例如在配对比对中加入空位)，这些失配不是被纠正而是被传播到最终结果。再者，更糟糕的是配对比对可能无法组成一个相容的多序列比对(图1-12)。以上因素使得渐进多序列比对对于距离非常接近的序列效果很好，而当序列间的距离较远时效果不佳。后期的渐进多序列比对软件对这些缺陷进行了改进。

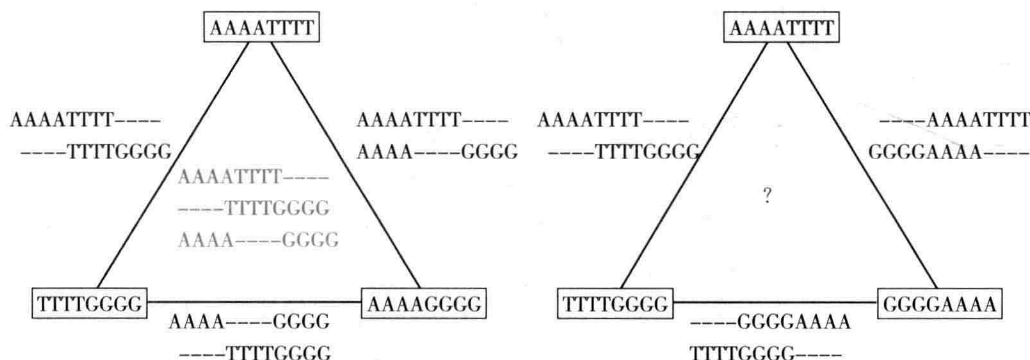


图1-12　三个序列的配对比对未必总能组合成一个多序列比对

(三) 迭代算法

在渐进多序列比对中，一个序列一经加入构造的比对结果其配对比对便不再重新处理，因此对在渐进比对过程中发现的错误或不适当的记分没有机会进行更正，这提高了比对的运行效率但牺牲了准确性。当起始的比对处理的是较远距离的序列时，其蕴含的错误对多序列比对的影响尤其严重。一类称作迭代法的方法能够克服渐进多序列比对的这个不足。迭代法的基本过程是先用渐进多序列比对产生一个初始结果，再对序列的不同子集进行反复比对并利用这些结果重新进行多序列比对，目标是改进多序列比对的总计分值。迭代法常常使用随机搜索或者通过对比对结果进行重排来寻找更优的解，迭代持续至比对记分值不再提高。

(四) 基于一致性的方法

渐进多序列比对的基本方法是先产生全部的配对比对，然后根据配对比对的计分高低逐渐构造多序列比对。基于一致性的方法采用了另一种利用序列信息的方式。这里，一致性指的是对于序列x、y和z，如果x_i比对于z_k且z_k比对于y_j，则x_i应比对于y_j。因此，基于一致性方法的基本特点是充分利用多个序列间的比对信息对配对比对进行更合理的计分。例如，根据x_i和y_j同时比对于z_k而调整x_i和y_j的比对计分，如果序列x中的字符x_i比对于序列y中的字符y_j的似然率(likelihood)为$P(x_i \sim y_j | x, y)$则有

$$P\left(x_i \sim y_j \mid x, y, z\right) \approx \sum_k P\left(x_i \sim z_k \mid x, z\right) P\left(y_j \sim z_k \mid y, z\right) \tag{1-3}$$

基于一致性的方法在多序列比对中对每对序列中的每对字符计算如上的似然率。根据基准测试数据的研究，基于一致性方法的多序列比对产生的结果经常比渐进多序列比对产生的结果更准确。

(五) 遗传算法

使用遗传算法的多序列比对把序列打碎成许多小片段，然后反复重组这些小片段，重组过程中通过在各个序列的不同位置引入空位来优化一个目标函数(通常是SP计分函数)，使得多个序列得以最优地比对。作为一种启发式算法，遗传算法不保证找到多序列比对的最优解，而且当超过20个序列时比对变得相当慢。一个用遗传算法对蛋白质序列进行比对的软件是SAGA(sequence alignment by genetic algorithm)。

(六) 模拟退火算法

模拟退火法的基本原理是，通过对一个由某个方法产生的多序列比对进行一系列重组而使比对进一步优化，因为这些重组有可能发现比原比对更优的比对。类似于遗传算法，模拟退火法也最大化一个类似于SP计分函数的目标函数，用于比对的定量评估。另外，它还使用一个“温度因子”(模拟退火法名称的由来)来决定重组的速率和每个重组发生的似然性。在典型的应用中，高重组率低似然性和低重组率高似然性交换使用，前者用于处理序列中的远距离区段而后者用于处理序列中的局部区段。一个使用模拟退火法的软件是MSASA(multiple sequence alignment by simulated annealing)。

(七) 隐马尔可夫模型

隐马尔可夫模型是一类概率模型,组成一个隐马尔可夫模型的要素包括: 一系列状态、每个状态间的转换概率、每个状态输出每个字符的概率以及由状态输出的字符所组成的序列。在基于隐马尔可夫模型的多序列比对中,DNA序列和蛋白质序列可看作是由不同的状态所产生的输出所构成的。当把一个碱基或一个氨基酸表示为一个节点并由此把要比对的多个序列用图表示时(这种图称有向无环图directed acyclic graph或偏序图partial-order graph,图1-13),多序列比对相当于对图进行简并,把每列中所有相同的字符归于一个节点中。特别是,若在一个列里所有的序列均有相同的字符,则它在有向无环图中仅被编码成一个节点; 若一个节点的下一列有n个不同的字符,则该节点有n个向外的导向这些字符的连接。另外,一个模型对空位、匹配和失配的每个可能组合都赋予一个概率。找出最优比对相当于找出最小公共超图(minimal common supergraph)(图1-13)。在这种隐马尔可夫模型里,可观测到的状态由一个个待比对序列的列所揭示,而隐含的状态表征了这些序列的祖先序列或共性序列。由于存在庞大的可能状态序列,一一搜索这些序列不切实际,求解这类隐马尔可夫模型的方法是所谓的Viterbi算法,它也是一个动态规划法的算法。基于不同的隐马尔科夫模型,人们开发了多个在计算效率和应用规模方面有所不同的软件,正确地使用不同的隐马尔可夫模型要比正确地使用不同的渐进多序列比对软件复杂。最简单的软件可能是POA(partial-order alignment),而一个风格类似但功

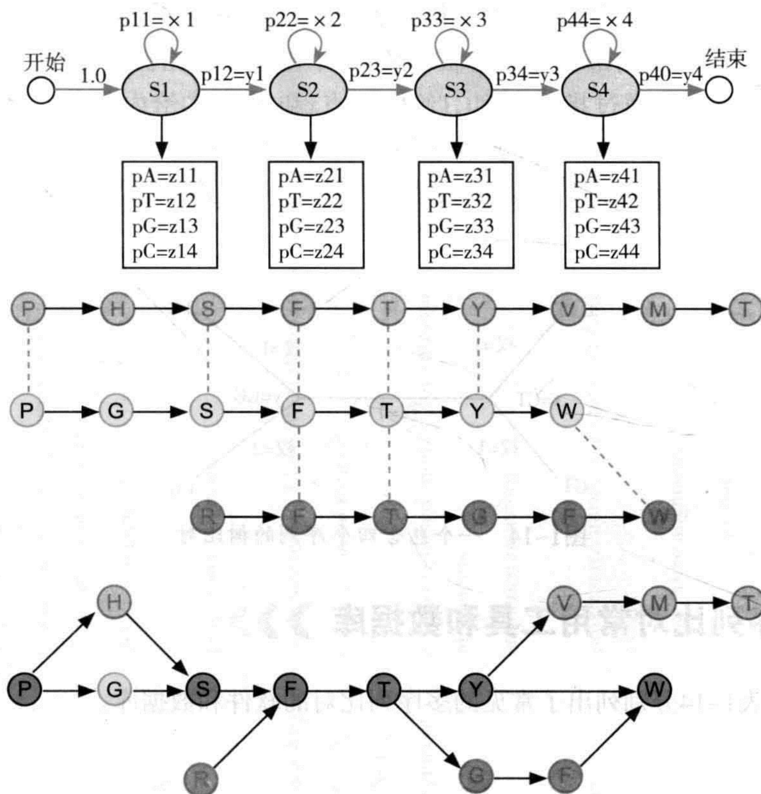


图1-13　隐马尔科夫模型和三个蛋白质序列PHSFTYVMT、PGSFTYW、RFTGFW的最小公共超图

能更广的软件是SAM(sequence alignment and modeling system),它被广泛用于蛋白质结构预测。使用隐马尔可夫模型进行多序列比对的长处包括它能对序列个数有较高要求。当序列间一致性较高时,需要20~50个序列进行多序列比对,而当序列间有较大变异时,可能需要多达100个序列来进行可靠的多序列比对。

(八) 星形比对和树形比对

星形比对是简单地基于一个固定序列与所有其他序列的配对比对而建立的,这个固定序列是星的中心。令$s_1,\cdots,s_k$是需比对的k个序列,为构造一个星比对,首先需挑选一个中心序列s_c,然后对每个下标不等于c的序列s_i使用动态规划法作s_c与s_i的双序列比对,费时$O(kn^2)$(假定序列长均为n)。接着遵循"一旦引入一个空位则始终保持这个空位"的原则将这些配对比对向s_c汇集,在此过程中不断地往s_c中加入空位以适配新加入的序列。中心序列的选择是星比对的关键,一个方法是逐个测试多个候选序列,择优而取,另一个方法是计算全部配对比对,然后选择使$\sum_{i\neq c} similarity(s_i, s_c)$最大的序列为中心序列。

当需要比对的序列可构成一棵进化树时,可以根据树的边所对应的配对比对计算全部序列整体的相似性,而不是用SP函数计算配对比对的相似性。具体方法是,假定有k个序列和一个恰有k个叶子的树,则树叶与序列具有一一对应性。如果对树的每个内部节点指派一个序列,就能计算每个边的权,它是与该边相连的两个节点所对应的两个序列间的相似性。树的计分,即全部序列整体的相似性,是所有边的权的和。树比对的任务是找出一个能使树记分最大的内部节点的指派(即为每个内部节点指派一个序列)。树比对的一个简单例子是图1-14,通过指派序列CT给内节点x和序列CG给内节点y,得到一个树计分6,计算该计分的规则是如果$a=b$则$p(a, b)=1$,否则为0;$p(a, -)=-1$。根据该规则,连接叶子序列CAT和内部序列$x=CT$的边的权值是1,而连接叶子序列CG和内部序列$y=CG$的边的权是2。

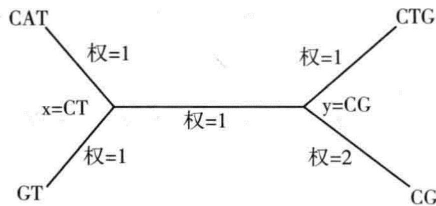


图1-14 一个包含四个序列的树比对

三、多序列比对常用工具和数据库

表1-13和表1-14分别列出了常见的多序列比对的软件和数据库。

表1-13 常见的多序列比对软件及其功能

名称	描述	序列类型	比对类型*	链接	完成时间
ABA	a-Bruijn alignment	Protein	Global	下载http: //nbcr.sdsc.edu/euler/aba_v1.0/	2004
AMAP	sequence annealing	Both	Global	在线http: //www.biomath.ucla.edu/msuchard/bali-phy	2006
BAli-Phy	tree+multi alignment; probabilistic/ Bayesian; joint estimation	Both	Global	在线+下载http: //www.biomath.ucla.edu/msuchard/ bali-phy	2005 2009
CHAOS/ DIALIGN	iterative alignment	Both	Local*	在线http: //dialign.gobics.de/chaos-dialign-submission	2003
ClustalW	progressive alignment	Both	Local or Global	在线http: //www.ebi.ac.uk/clustalw/ 下载http: //www.clustal.org/	1994
Codon Code Aligner	multi alignment; ClustalW & Phrap support	Nucleotides	Local or Global	下载http: //www.codoncode.com/aligner/	2003 2009
DIALIGN-TX, DIALIGN-T	segment-based method	Both	Local* Global	下载+在线http: //dialign-tx.gobics.de/	2005 2008
DNA Alignment	segment-based method for intraspecific alignments	Both	Local* or Global	在线http: //www.fluxus-engineering.com/align.htm	2005 2008
FSA	sequence annealing	Both	Global	下载http: //fsa.sourceforge.net/index.html 在线http: //orangutan.math.berkeley.edu/fsa/	2008
Geneious	progressive/iterative alignment; ClustalW plugin	Both	Local or Global	下载http: //www.geneious.com/	2005 2009
Kalign	progressive alignment	Both	Global	在线http: //www.ebi.ac.uk/kalign/	2005
MSA	dynamic programming	Both	Local or Global	下载http: //www.ncbi.nlm.nih.gov/CBBresearch/ Schaffer/msa.html	1989 1995
PRRN/ PRRP	iterative alignment (especially refinement)	Protein	Local or Global	在线http: //align.genome.jp/prrn/	1991
POA	partial order/hidden Markov model	Protein	Local or Global	下载+在线http: //bioinfo.mbi.ucla.edu/poa/	2002

续表

名称	描述	序列类型	比对类型*	链接	完成时间
SAM	hidden Markov model	Protein	Local or Global	在线http: //compbio.soe.ucsc.edu/sam.html	1994 2002
MARNA	multiple alignment of RNAs	RNA	Local	在线http: //biwww2.informatik.uni-freiburg.de/Software/MARNA/index.html 下载http: //biwww2.informatik.uni-freiburg.de/Software/MARNA/download/index.html	2005
MAFFT	progressive/iterative alignment	Both	Local or Global	在线http: //align.bmr.kyushu-u.ac.jp/mafft/online/server/	2005
MAVID	progressive alignment	Both	Global	在线http: //baboon.math.berkeley.edu/mavid/	2004
MULTALIN	dynamic programming/clustering	Both	Local or Global	在线http: //prodes.toulouse.inra.fr/multalin/multalin.html	1988
Multi-LAGAN	progressive dynamic programming alignment	Both	Global	在线http: //genome.lbl.gov/vista/lagan/submit.shtml	2003
MUSCLE	progressive/iterative alignment	Both	Local or Global	在线http: //www.drive5.com/muscle	2004
Pecan	probabilistic/consistency	DNA	Global	下载http: //www.ebi.ac.uk/~bjp/pecan/	2008
ProbCons	probabilistic/consistency	Protein	Local or Global	在线http: //probcons.stanford.edu/index.html	2005
PSAlign	alignment preserving non-heuristic	Both	Local or Global	下载http: //faculty.cs.tamu.edu/shsze/psalign/	2006
SAGA	sequence alignment by genetic algorithm	Protein	Local or Global	下载http: //www.tcoffee.org/Projects_home_page/saga_home_page.html	1996 1998
StatAlign	Bayesian co-estimation of alignment and phylogeny	Both	Global	下载http: //phylogeny-cafe.elte.hu/StatAlign/	2008
T-Coffee	more sensitive progressive alignment	Both	Local or Global	在线http: //www.tcoffee.org/ 下载http: //www.tcoffee.org/Projects_home_page/t_coffee_home_page.html	2000 2008

注：*指的是主要用于这种比对

表1-14　常见的多序列比对数据库

数据库名称	描述	网址
BLOCKS	类隐与模型库，无空隙	http: //blocks.fhcrc.org/
InterPro	联合了PROSITE、PRINTS、ProDom Pfam、SMART、TIGRfam的资源	http: //www.ebi.ac.uk/interpro/
CDD	保守结构域数据库	http: //www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
Pfam	隐马模型库	http: //pfam.sanger.ac.uk/
PRINTS	SwissProt/TrEMBL的蛋白指纹	http: //bioinf.man.ac.uk/dbbrowser/PRINTS/index.php
PROSITE	蛋白质模体字典	http: //prosite.expasy.org/

(一) CLUSTAL W

CLUSTAL W软件是一个目前最为普遍使用的多序列比对程序(http: //www.ebi.ac.uk/Tools/msa/clustalw2/)，采用渐进的多序列比对方法，先将多个序列两两比对构建距离矩阵，反映序列之间两两关系；然后根据距离矩阵计算产生系统进化指导树，对关系密切的序列进行加权；然后从最紧密的两条序列开始，逐步引入邻近的序列并不断重新构建比对，直到所有序列都被加入为止。CLUSTAL W程序有很多版本，可以基于UNIX、DOS和WINDOWS等多种操作平台同时被许多常用的序列分析软件所集成。从ftp: //ftp.ebi.ac.uk/pub/software/clustalw2/地址可以得到它的不同版本。ClustalW的在线服务界面见图1-15，目前它的最高版

ClustalW2 - Multiple Sequence Alignment
ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins.
New version! Clustal Omega is now available for protein sequences - give it a try!
Use this tool
STEP 1 - Enter your input sequences
Enter or paste a set of Protein sequences in any supported format:
Or, upload a file: 选择文件 未选择文件
STEP 2 - Set your Pairwise Alignment Options
Alignment Type: Slow Fast
The default settings will fulfill the needs of most users and, for that reason, are not visible.
More options... (Click here, if you want to view or change the default settings.)
STEP 3 - Set your Multiple Sequence Alignment Options
The default settings will fulfill the needs of most users and, for that reason, are not visible.
More options... (Click here, if you want to view or change the default settings.)

图1-15　ClustalW的在线服务界面

本是2.1。对于大多数使用者来说，普遍采用的是运行于WINDOWS界面的版本CLUSTAL X。目前，该程序的最新版本是CLUSTAL X 2.1。该程序能支持7种输入序列格式。当序列中超过85%是A，C，G，T，U，N时，程序自动认为这是个核酸序列，否则认为是蛋白质序列，通常使用者采用Fasta格式。例如，通过NCBI的Entrez工具检索分别得到人（*Homo sapiens*）、小鼠（*Mus musculus*）、大鼠（*Rattus norvegicus*）和鸡（*Gallus gallus*）等四种动物的发状分裂相关增强子-5蛋白序列，将这些序列以FASTA格式投入到CLUSTAL W界面输入框中，点击，得到多序列比对结果见图1-16。

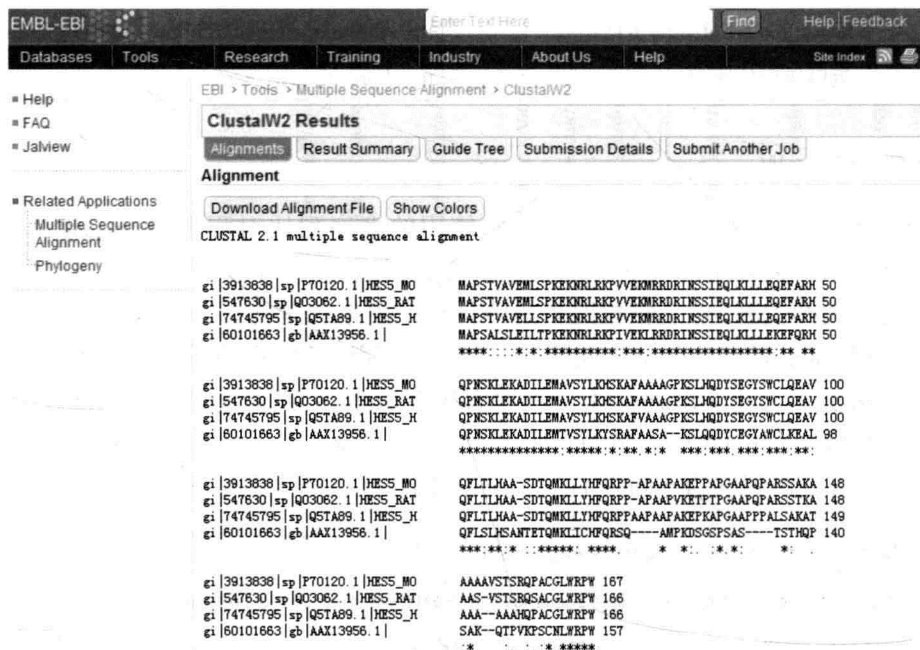


图1-16 ClustalW的多序列比对结果

(二) MUSCLE

自2004年，MUSCLE（multiple sequence alignment by log-expectation，http: //www.ebi.ac.uk/Tools/msa/muscle/）由于其准确性和出色的速度而成为一个流行的用于大量序列多序列比对的软件。MUSCLE的在线服务界面如图1-17所示。据报道，使用桌面计算机MUSCLE可以在21秒内完成1000个长度为282的蛋白质序列的比对。MUSCLE的方法分为两个步骤。首先，使用渐进多序列比对产生一个初始结果，其中含有根据每对序列的相似性计分构造的一棵指导树。其次，重新计算相似性计分，据此改进指导树并再用渐进多序列比对产生一个更新的结果。这一过程迭代地进行，而算法根据新计算的SP计分值是否增加而决定是接受还是拒绝新产生的比对结果。

(三) ProbCons

基于一致性的多序列比对软件ProbCons（probabilistic consistency-based multiple alignment，http: //probcons.stanford.edu/）。ProbCons的在线服务界面见图1-18。ProbCons分五步进行蛋白质多序列比对。第一，对每对序列中的每对字符计算似然率，得到一个似然率矩阵。第二，

MUSCLE - Multiple Sequence Alignment

MUSCLE stands for **MU**ltiple **S**equence **C**omparison by **L**og- **E**xpectation. MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee, depending on the chosen options.

Use this tool

STEP 1 - Enter your input sequences
Enter or paste a set of sequences in any supported format:

Or upload a file: 选择文件 未选择文件

STEP 2 - Set your Parameters
OUTPUT FORMAT: Pearson/FASTA

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 3 - Submit your job
☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

图1-17　MUSCLE在线服务界面

PROBCONS

Probabilistic Consistency-based Multiple Alignment of Amino Acid Sequences

RUN　ABOUT　DOWNLOAD　HELP

PROBCONS is an efficient protein multiple sequence alignment program, which has demonstrated a statistically significant improvement in accuracy compared to several leading alignment tools.

The email server is currently down. In the meantime, please try the CONTRAlign server web interface.

BASIC PARAMETERS

E-mail address

E-mail address (again)

Input sequence file　选择文件 未选择文件

ADDITIONAL OPTIONS

Consistency reps　2

Iterative refinement reps　100

Pre-training reps　0

Output format　○ MFA　◉ CLUSTALW

COMPUTE ALIGNMENT

图1-18　ProbCons的在线服务界面

用动态规划法计算每个配对比对的预期精度(expected accuracy),它是得到正确比对的字符数除以较短序列的长度,计分根据上述条件概率公式计算而不采用通常的PAM或BLOSUM矩阵,且空位罚分设为0。第三,根据相关条件概率的计算重新调整配对比对的计分,这一步用到了由多个配对比对揭示的序列中字符的保守性,产生更准确的对替换的记分。第四,用分层聚类法(hierarchical clustering)构造一棵基于相似性而不是距离的期望准确性指导树。第五,根据该期望准确性指导树对所有的序列进行渐进性比对,方法如同ClusterW。在这些步骤之后,还可进一步用迭代法进行优化。

(四) MultAlin

MultAlin是一个基于Web服务的程序,可登录http: //www-archbac.u-psud.fr/genomics/multalin.html上执行多序列比对。MultAlin方法也是从一系列的两两比对开始,计算出相似性分值,再根据这些分数值进行分层聚类。当序列都被分类后,进行多序列比对,计算出多序列比对中序列两两比对的新数值,基于这些数值,再做新分类,这个过程不端循环,直到相似性分数值不再上升为止。

(五) Pfam

Pfam是一个综合的蛋白质家族的大集合,同时收集了序列多重比对和蛋白质家族的profile HMMS。在Pfam数据库中可以选择蛋白质及DNA序列搜索,关键词搜索,也可以选择查看Pfam数据库的多序列比对信息(BROWSE PFAM),以及分类搜索(TAXONOMY SERCH),还可以看到关于Pfam的帮助信息。

Pfam数据库由两个部分组成: Pfam-A和Pfam-B。Pfam -A的质量比较高,是手工编辑、多重比对格式的蛋白质家族集合。对于每一个家族,Pfam提供了4种特征:注释、种子比对、profile HMM和完全比对。完全比对可能很大,Pfam前20个家族的完全比对都含有超过2500个序列。种子比对含有较少数量的代表序列。虽然这些Pfam -A的数据涵盖了在许多基础序列数据库中很大的比例,为了让更多的全面了解已知蛋白质,另外一些从ProDom数据库自动生成的被称为 Pfam - B。虽然质量较低,Pfam - B可以被用来鉴别功能保守区域,尤其是没有Pfam -A的时候。

由于存在众多的多序列比对方法和软件,选择合适的软件既十分重要又常常不易。可遵循如下几条原则。首先,序列的种类影响软件的选择。有些软件专用于蛋白质或DNA序列,有些软件则两者皆可。比对蛋白质、cDNA和RNA序列时一般选择全局比对,因为整个序列常常是一个功能单元,而比对DNA序列时应考虑glocal或syntenic比对,因为DNA序列中常常同时包含保守和非保守的区段。其次,比对的目的影响软件的选择。如果蛋白质和RNA序列可能包含多个保守的域(domain),且比对的目的是发现这些域,则应选用syntenic比对。发现多个域的典型情形是寻找一个基因中被多个内含子分隔的多个外显子、一个蛋白质中被多个非保守域分隔的多个保守域和一段基因调节区中被多个非保守区段分隔的保守位点。第三,序列的长短影响软件的选择。MSA不能比对超过500字符的序列,比对较长的DNA序列可用MAP2,而比对整条染色体甚至整个基因组时通常使用UCSC基因组浏览器和Ensembl基因组浏览器。第四,序列保守性的程度可影响软件的选用。在许多DNA序列中,保守区段的保守性介乎于高度保守的外显子和完全不保守的junk DNA之间,不易由常规的记分机制

得以揭示，而UCSC基因组浏览器中的phastCons和phyloP提供了可有效揭示这种中度保守性的计算方法。第五，种系关系的距离影响软件的选择。当序列间种系距离较近时，许多软件会产生大致相同的结果，反之，当序列间种系距离较远时，不同软件产生的结果可能会有相当大的差异，使用基于一致性的方法可充分利用序列间的种系信息。另外，对于比对远距离种系的序列，对敏感性和选择性的取舍十分重要。敏感性关乎识别尽可能多的同源区段，选择性要求识别的同源区段都是真的。不同的软件在这两个彼此矛盾的指标上有不同的取舍。第六，比对种系关系已知的序列时，可使用利用指导树或种系树的算法和软件，比对种系关系未知的序列时，则无法使用这样的软件。对于全基因组序列比对是否使用参照序列以及选用什么序列作为参照序列，这取决于具体序列的特征（包括序列间距离的远近）、对序列的了解（包括对参照序列的了解）、比对的目的（是否主要揭示直系同源区段）以及对比对质量的预估。第七，因为不同算法具有不同的时间和空间复杂度，序列的数量、长度和计算机的性能也影响实际算法和软件的选用。

第四节 核酸序列特征分析

Section 4 Analysis of DNA Sequence Characteristics

一、基因开放读码框的识别

开放读码框(Open Reading Frame, ORF)是DNA上的一段碱基序列,包括从5'端翻译起始密码子(ATG)到终止密码子(TAA、TAG、TGA)的编码蛋白质的碱基序列。每个ORF对应一个潜在的蛋白质编码区域。对于任意给定的一段DNA序列,我们并不知道DNA双链中哪一条是编码链,也不能确定其编码区是否从这条序列的第一个碱基开始所以每条链都有3种潜在的开放读码框,一段双链DNA序列在理论上就有6种潜在的开放读码框,即先以所给的DNA单链为模板,分别从5'→3'方向的第1、2、3个碱基开始翻译,再以其互补链为模板,分别从3'→5'方向的第1、2、3个碱基开始翻译,得到另外3种翻译结果。正链上的3个读码框称为"正向"(forward)读码框,而负链(或互补链)上的读码框称为"反向"(reverse)读码框。在6个潜在的开放读码框中,一般选择中间没有被终止密码子隔开最大的那个读码框作为正确的预测结果。

原核生物的基因结构比较简单,绝大多数是连续基因,不含间隔的内含子。多数基因组的编码序列都在100个氨基酸以上。真核生物的基因结构远比原核生物的复杂。真核生物的基因一般为断裂基因(interrupted gene),由内含子和外显子组成,编码区被内含子分隔成若干段,开放读码框的长度变化范围非常大,因此真核生物基因结构的预测远比原核生物困难。但是,在真核生物的开放读码框中,外显子与内含子之间的连接在绝大部分情况下满足GU-AG规律:内含子序列 5'端起始的两个核苷酸总是GU,并且其3'端最后的两个核苷酸总是AG,即:5'-GU ……AG-3',这个规律有助于真核生物开放阅读框的识别。

目前国际上用于开放读码框的预测工具有很多(表1-15),这些工具使用的预测方法、针对的物种范围和最终的结果都各有不同。这些预测工具按照预测方法的不同主要分为两类:第一种方法以统计学分析和模式识别为基础(statistics-based)的方法,从基因序列本身进行预测,不需要与大规模的数据库进行比较,预测速度快,当缺少待分析物种的相关数据库信息时用这种方法是比较好的选择,GENSCAN就是基于这种方法建立的工具,使用比较广泛,预测效率比较高。第二种方法是以同源比对为基础(homology-based)的方法,依赖于已知的数据库来源、数量和质量,预测的正确性比第一类高。以人发状分裂相关增强子-5的mRNA序列和ORF Finder工具为例,其在GenBank中的编码为BC087840。从GenBank中下

载此序列并粘贴到ORF Finder指定的框内(图1-19),点击"OrfFind"按钮提交序列,六框翻译的结果见图1-20,其中通常只有一条是可读框,一般很难随机发现很长的ORF,因而长的ORF很可能意味着存在CDS。

表1-15　开放读码框识别常用相关工具列表

工具名	网址	研发者	适用范围
ORF Finder	http: //www.ncbi.nlm.nih.gov/gorf/gorf.html	NCBI	通用
BESTORF	http: //linux1.softberry.com/berry.phtml ? topic=bestorf&group=programs&subgroup=gfind	Softberry	真核
GENSCAN	http: //genes.mit.edu/GENSCAN.html	MIT	脊椎、拟南芥、玉米
GlimmerM	http: //cbcb.umd.edu/software/glimmer/	Maryland	原核
Gene Finder	http: //rulai.cshl.org/tools/genefinder/	Zhang' s Lab	人、小鼠、拟南芥、酵母
GeneMark	http: //opal.biology.gatech.edu/GeneMark/	GIT	通用

图1-19　ORF Finder的在线操作界面

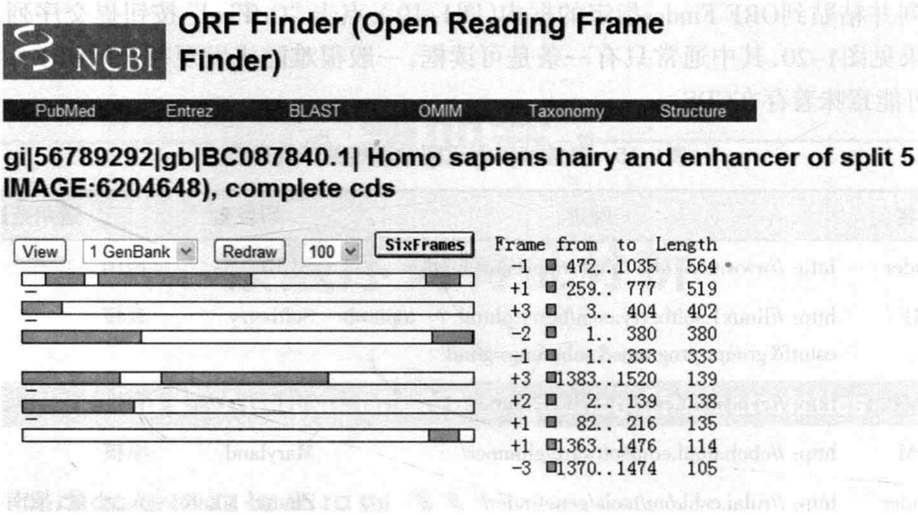


图1-20 ORF Finder的六框翻译结果

二、内含子/外显子剪切位点的识别

真核生物的基因一般为断裂基因(interrupted gene),由内含子和外显子组成,编码序列通常被内含子隔开。虽然内含子的长度没有一定的规律,但是内含子和外显子的边界和周围序列是由前体mRNA内的具有保守性的一些特殊核苷酸序列表明的,通常内含子5’端剪切位点以GU开始,称为供体位点(donor),3’端剪切位点以AG结束,称为受体位点(acceptor),还包括一个位于内含子内,靠近3’端的分支位点(常为A),后面为多聚嘧啶区。在分析基因组数据时,经常需要预测基因的RNA选择性剪切方式,即内含子和外显子的位置和数量。预测是基于RNA剪接的保守性序列“GU-AG”规则。根据这一特点并结合ORF, Blast等数据就可以对未知基因的成熟mRNA序列进行预测。表1-16列出了一些常见的内含子/外显子剪切位点识别工具。一般来说基因组核苷酸序列的包含剪切位点和内含子可用NetGene2和Splice View等工具直接预测;而对于mRNA/cDNA序列的分析,则需要借助Spidey, SIM4, BLAT和BLAST等序列比对工具从相应的基因组序列推断基因结构。

表1-16 常见的内含子/外显子剪切位点识别工具

工具名	网址	研发者	适用范围
NetGene2	http: //www.cbs.dtu.dk/services/NetGene2/	CBS	人类、线虫、拟南芥
GeneSplicer	http: //cbcb.umd.edu/software/GeneSplicer/	CBCB	恶性疟原虫,人类、拟南芥、果蝇、水稻
Spidey	http: //www.ncbi.nlm.nih.gov/spidey/	NCBI	脊椎、果蝇、线虫、植物
GeneSeqer	http: //deepc2.psi.iastate.edu/cgi-bin/gs.cgi	ISU	通用

三、序列模体的查找和可视化工具

模体(Motif)是指序列中局部的保守区域,或者是一组序列中共有的一小段序列模式。

更多的时候是指有可能具有分子功能、结构性质或家族成员相关的任何序列模式。MEME软件包(http: //meme.sdsc.edu/meme/intro.html)是一个对DNA序列或者蛋白质序列中模体进行识别和分析的一个综合性的工具。表1-17列出了MEME软件包中提供的满足各种不同需要的工具。

表1-17　MEME软件包中各种工具

工具名	网址	功能
MEME	http: //meme.nbcr.net/meme/cgi-bin/meme.cgi	模体识别
GLAM2	http: //meme.nbcr.net/meme/cgi-bin/glam2.cgi	模体识别
MEME-ChIP	http: //meme.nbcr.net/meme/cgi-bin/meme-chip.cgi	模体识别
FIMO	http: //meme.nbcr.net/meme/cgi-bin/fimo.cgi	模体搜索
GLAM2SCAN	http: //meme.nbcr.net/meme/cgi-bin/glam2scan.cgi	模体搜索
MAST	http: //meme.nbcr.net/meme/cgi-bin/mast.cgi	模体搜索
SPAMO	http: //meme.nbcr.net/meme/cgi-bin/spamo.cgi	模体间距分析
MCAST	http: //meme.nbcr.net/meme/cgi-bin/mcast.cgi	模体簇搜索
TOMTOM	http: //meme.nbcr.net/meme/cgi-bin/tomtom.cgi	模体比较
GOMO	http: //meme.nbcr.net/meme/cgi-bin/gomo.cgi	模体功能分析

四、密码子使用模式的分析

由于密码子简并性的存在,每个氨基酸至少对应一种密码子,最多有6种对应的密码子。编码同一种氨基酸的密码子称为同义密码子。不同物种、不同基因在密码子使用上都存在着很大的差异。各种生物体似乎更偏爱使用某些同义三联密码子。例如,某一物种或基因通常倾向于使用一种或者几种特定的同义密码子,这些密码子被称为最优密码子,此现象被称为密码子使用偏好性。密码子使用偏好性的产生与基因的表达水平、翻译起始效应、基因的碱基组成、GC含量、基因长度,tRNA的丰度等很多因素相关。密码子分析常用软件和常用网站见表1-18和表1-19。

表1-18　密码子使用偏好性分析常用软件

软件	网址	操作系统
CodonW	http: //sourceforge.net/projects/codonw/	Dos, unix, windows
SYCO	http: //emboss.sourceforge.net/apps/cvs/emboss/apps/syco.html	Unix, linux
CHIPS	http: //www.cbib.u-bordeaux2.fr/pise/chips.html	Unix, linux
CUSP	http: //emboss.sourceforge.net/apps/cvs/emboss/apps/cusp.html	Unix, linux
CodonPreference	http: //odin.mdacc.tmc.edu/gcg/unix/codonpreference.html	Unix, linux
CodonFrequency	http: //bioinfo.ekmd.huji.ac.il/gcg11manual/codonfrequency.html	Unix, linux
Correspond	http: //bioinfo.ekmd.huji.ac.il/gcg11manual/correspond.html	Unix, linux
Countcodon	http: //www.kazusa.or.jp/codon/countcodon.html	Web

表1-19 密码子使用偏好性分析常用网站

网址	特点
ftp: //ftp.kazusa.or.jp/pub/codon/current/ ftp: //ftp.ebi.ac.uk/pub/databases/cutg/ ftp: //ftp.nig.ac.jp/pub/db/codon/current/	CUTG,密码子使用频度表。由 GenBank 中的 DNA 序列统计出来的密码子使用频度表(Codon Usage Tabulate from GenBank),按物种和模式生物给出。把蛋白质氨基酸序列倒翻译为核苷酸序列时,应参考此表
http: //www.kazusa.or.jp/codon/	CUTG(Codon Usage Tabulated from GenBank)的网络扩展版,可以查询不同物种的密码子使用表
http: //gcua.schoedl.de/	以图形的方式表现密码子偏好性
http: //bioinformatics.org/codon/cgi-bin/codon.cgi	将 Codon Usage Database 中的所关心物种的密码子表,经处理转化为可读性更强的图表形式
http: //www.faculty.ucr.edu/~mmaduro/codonusage/usage.htm	它对于在 E.coli 中异源蛋白的表达效率给出了很好的建议

注: 来源于吴宪明,吴松峰,任大明,等.密码子偏性的分析方法及相关研究进展.遗传.2007;29(4):420-426。

五、限制性核酸内切酶位点分析 》》》

限制性核酸内切酶(以下简称限制性酶)是一类识别双链DNA中特定核苷酸序列的DNA水解酶,以内切方式水解DNA,产生5'-P和3'-OH末端。限制性酶的识别序列,大部分具有双轴对称性结构或称回文序列,具有一定的保守性,利用这一特性可以识别基因序列中的限制性核酸内切酶位点。表1-20列出了常用的限制性核酸内切酶位点分析工具。

Vector NTI软件输入文件格式广泛,除了molecule documents(.gb)是该公司本身文件格式外,还能识别各种数据库应用格式软件: EMBL、GenBank、FASTA、Sequence files。可以查找特定序列、ORF(可以设置相关参数)、描述载体、限制酶位点、一些功能序列和附注。整个界面由文本、图形和序列三部分构成,而且点击任意的序列、RE、基因、图形和序列均会自动标记到相应位置,非常直观方便。载体可以圆形表示也可以线形表示。还可进行核酸到蛋白的翻译等功能。

表1-20 常用的限制性核酸内切酶位点分析工具

工具	网址	备注
Vector NTI	http: //register.informaxinc.com/solutions/vectornti/index.html	Windows
Webcutter	http: //bio.lundberg.gu.se/cutter2/	Web
Watcut	http: //watcut.uwaterloo.ca/watcut/watcut/template.php	Web
NEBcutter	http: //tools.neb.com/NEBcutter2/index.php	Web
BioEdit	http: //www.mbio.ncsu.edu/BioEdit/bioedit.html	Windows
DNAMAN	http: //www.lynnon.com/	Windows
RestrictionMapper	http: //www.restrictionmapper.org/	Web

六、重复序列的查找

重复序列(repetitive sequence)是指真核生物染色体基因组中重复出现的核苷酸序列。这些序列一般不编码多肽,其组织形式有两种: 串联重复序列; 分散重复序列。前一种成簇存在于染色体的特定区域,后一种分散于染色体的各位点上。重复DNA序列是多数真核生物基因组的主要成分,可以分为三个主要类型: 低重复序列、中度重复序列和高度重复序列。重复序列中往往GC含量低,AT含量高,3’端和5’端有直接重复序列的存在。有利于形成环状结构。对这些重复序列的定位能为基因定位提供重要的反向信息,同时重复序列还常会干扰序列其他特性分析。表1-21列出了常见的重复序列查找工具。

表1-21　常见的重复序列查找工具

工具	网址	备注
REPFIND	http: //zlab.bu.edu/repfind/	Web
RepeatMasker	http: //www.repeatmasker.org/	Web, linux

第五节
蛋白质序列特征分析

Section 5 Analysis of Protein Sequence Characteristics

一、蛋白质的理化性质分析

蛋白质理化性质是蛋白质研究的基础，对组成蛋白质的氨基酸进行理化性质的统计分析是对未知蛋白质进行分析的基础。蛋白质的理化性质包括相对分子质量、氨基酸组成、等电点、消光系数、半衰期、不稳定系数和总平均亲水性等。传统的理化性质分析方法如相对分子质量的测定、等电点实验和沉降实验等十分费时和耗资。基于实验经验值的计算机分析方法为蛋白质的理化性质分析提供了一个便捷的途径。表1-22列出了一些常用的蛋白质理化性质分析工具。

表1-22　蛋白质理化性质常用分析工具

工具	网址	备注
ProtParam	http: //us.expasy.org/tools/protparam.html	Web
ProtScale	http: //ca.expasy.org/tools/protscale.html	Web
Compute pI/MW	www.expasy.ch/tools/	Web
TGREASE	ftp: //ftp.virginia.edu/pub/fasta/	Windows
SAPS	www.isrec.isb-sib.ch/software/SAPS_form.html	Web

ExPASy(expert protein analysis system)是由瑞士生物信息学中心维护，并与欧洲生物信息学中心(EBI)及蛋白质信息资源(protein in formation resource, PIR)组成Universal Protein Knowledgebase(Uniprot)联盟。ExPASy数据库提供了一系列蛋白质理化分析工具，以便于检索未知蛋白质的理化性质，并基于这些理化性质鉴别未知蛋白质的类别，为后续实验提供帮助。其中ProtParam(physico-chemical parameters of a protein sequence)就是计算氨基酸理化参数常用的在线工具，其网址为http: //expasy.org/tools/protparam.html, ProtParam提供的理化性质主要包括氨基酸残基数(number of amino acids)、分子质量(molecular weight)、理论等电点(theoretical pI)、氨基酸组成(amino acid composition)、负电荷氨基酸残基总数(total number of negatively charged residues)、正电荷氨基酸残基总数(total number of positively charged residues)、原子组成(atomic composition)、分子式(formula)、原子总数(total number of atoms)、消光系数(extinction

coefficients)、半衰期(estimated half-life)、不稳定系数(instability index)、脂肪系数(aliphatic index)、总平均疏水性(grand average of hydropathicity)等物理和化学参数。ExPASy的ProtScale程序是计算蛋白质亲疏水性分析的在线工具,其网址为http: //expasy.org/tools/protscale.html,用于计算氨基酸标度(amino acid scale)。氨基酸标度表示氨基酸在某种实验状态下相对其他氨基酸在某些性质的差异,如疏水性、亲水性等。ProtScale程序收集了50多个文献中提供的氨基酸标度,默认值为Hphob.Kyte & Doolittle,做疏水性分析,可以对一些处于蛋白质分子表面的抗原决定簇及一些膜蛋白中穿越膜的肽段进行预测。以人发状分裂相关增强子-5的蛋白质为例,其在GenBank中的编码为Q5TA89。从GenBank中下载此序列并粘贴到ProtParam指定的框内见图1-21,点击Compute parameters按钮提交序列,蛋白质序列的理化性质分析结果见图1-22。

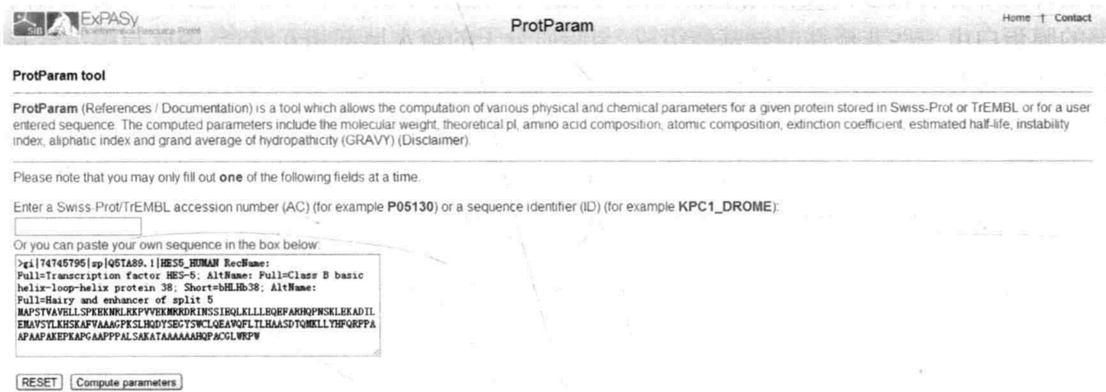


图1-21 ProtParam的在线操作界面

```
Carbon      C      818
Hydrogen    H     1299
Nitrogen    N      231
Oxygen      O      229
Sulfur      S        6
```

Formula: $C_{818}H_{1299}N_{231}O_{229}S_6$
Total number of atoms: 2583

Extinction coefficients:

Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.

```
Ext. coefficient     22585
Abs 0.1% (=1 g/l)    1.239, assuming all pairs of Cys residues form cystines


Ext. coefficient     22460
Abs 0.1% (=1 g/l)    1.232, assuming all Cys residues are reduced
```

Estimated half-life:

The N-terminal of the sequence considered is M (Met).

```
The estimated half-life is: 30 hours (mammalian reticulocytes, in vitro).
                            >20 hours (yeast, in vivo).
                            >10 hours (Escherichia coli, in vivo).
```

Instability index:

The instability index (II) is computed to be 56.31
This classifies the protein as unstable.

Aliphatic index: 79.64

图1-22 ProtParam分析蛋白质序列理化性质的部分结果

二、蛋白质的跨膜结构分析

生物膜所含的蛋白质叫膜蛋白，是生物膜功能的主要承担者。根据蛋白质分离的难易及在膜中分布的位置，膜蛋白基本可分为两大类：外在膜蛋白和内在膜蛋白。外在膜蛋白约占膜蛋白的20%~30%，分布在膜的内外表面，主要在内表面，为水溶性蛋白质，它通过离子键、氢键与膜脂分子的极性头部相结合，或通过与内在蛋白的相互作用间接与膜结合；内在蛋白约占膜蛋白的70%~80%，是双亲媒性分子，可不同程度的嵌入脂双层分子中。有的贯穿整个脂双层，两端暴露于膜的内外表面，这种类型的膜蛋白又称跨膜蛋白。内在膜蛋白露出膜外的部分含较多的极性氨基酸，属亲水性，与磷脂分子的亲水头部邻近；嵌入脂双层内部的膜蛋白由一些非极性的氨基酸组成，与脂质分子的疏水尾部相互结合，因此与膜结合非常紧密。含有跨膜区的蛋白质往往和细胞的功能状态密切相关。表1-23列出了跨膜结构分析常用的工具。

表1-23 蛋白质跨膜结构分析常用的工具

工具	网址	备注
Tmpred	http: //www.ch.embnet.org/software/TMPRED_form.html	Web
TMHMM	http: //www.cbs.dtu.dk/services/TMHMM/	Web
PSORT	http: //psort.hgc.jp/form.html	Web
DAS	http: //www.sbc.su.se/~miklos/DAS/	Web
SPLIT	http: //split.pmfst.hr/split/	Web
PRED-TMR	http: //athina.biol.uoa.gr/PRED-TMR/	Web

TMpred是EMBnet开发的分析蛋白质跨膜区的在线工具，其网址为http: //www.ch.embnet.org/software/TMPRED_form.html 。TMpred基于对TMbase数据库的统计分析来预测蛋白质跨膜区和跨膜方向。TMbase来源于Swiss-Prot库，并包含了每个序列的一些附加信息，如：跨膜结构区域的数量、跨膜结构域的位置及其侧翼序列的情况。Tmpred利用这些信息并与若干加权矩阵结合进行预测。用户将一个蛋白质序列输入查询序列文本框，并可以指定预测时采用的跨膜螺旋疏水区的最小长度和最大长度。输出结果包含四个部分：可能的跨膜螺旋区、相关性列表、建议的跨膜拓扑模型以及表示相同结果的图。以G蛋白偶联受体蛋白质序列为例，其在GenBank中的编号为P51684。将“P51684”输入到TMpred的查询序列文本框中，输入序列格式选择“SwissProt ID or AC”见图1-23，按“Run TMpred”按钮，可得到TMpred对P51684序列的分析结果。图1-24到图1-27显示了用TMpred分析P51684序列得到的7个可能的跨膜螺旋区、7个跨膜螺旋区的相关性列表、建议的跨膜拓扑模型和图形显示结果。

ch.EMBnet.org

Home　Services　Courses　Links　Contacts

TMpred - Prediction of Transmembrane Regions and Orientation

The TMpred program makes a prediction of membrane-spanning regions and their orientation. The algorithm is based on the statistical analysis of TMbase, a database of naturally occuring transmembrane proteins. The prediction is made using a combination of several weight-matrices for scoring.

K. Hofmann & W. Stoffel (1993)
TMbase - A database of membrane spanning proteins segments
Biol. Chem. Hoppe-Seyler **374**,166

For further information see the TMbase and TMpredict documentation.

Usage: Paste your sequence in one of the supported formats into the sequence field below
and press the "Run TMpred" button.
Make sure that the format button (next to the sequence field) shows the correct format
Choose the minimal and maximal length of the hydrophic part of the transmembrane helix

Output format	html　minimum 17　maximum 33
Query title (optional)	
Input sequence format	SwissProt ID or AC
Query sequence: or ID or AC or GI (see above for valid formats)	P51684
	Run TMpred　Clear Input

图1-23　TMpred的在线操作界面

1.) Possible transmembrane helices

The sequence positions in brackets denominate the core region.
Only scores above 500 are considered significant.

Inside to outside helices : 7 found

from		to		score	center
47	(51)	69	(69)	2494	61
83	(86)	104	(104)	1914	94
123	(123)	141	(139)	1352	131
166	(168)	184	(184)	2170	176
219	(219)	236	(236)	2453	227
255	(255)	276	(273)	2140	265
300	(300)	319	(319)	915	309

Outside to inside helices : 7 found

from		to		score	center
55	(55)	74	(71)	2707	63
84	(86)	104	(104)	1470	94
120	(123)	141	(139)	1451	131
166	(166)	185	(185)	1934	176
212	(214)	235	(232)	2530	224
252	(258)	274	(274)	1386	266
299	(299)	319	(319)	1299	309

图1-24　用TMpred分析P51684序列所得到的7个可能跨膜螺旋区

2.) Table of correspondences

Here is shown, which of the inside->outside helices correspond to which of the outside->inside helices.

Helices shown in brackets are considered insignificant.
A "+"-symbol indicates a preference of this orientation.
A "++"-symbol indicates a strong preference of this orientation.

```
        inside->outside  | outside->inside
   47-  69 (23) 2494      |    55-  74 (20) 2707 ++
   83- 104 (22) 1914 ++   |    84- 104 (21) 1470
  123- 141 (19) 1352      |   120- 141 (22) 1451  +
  166- 184 (19) 2170 ++   |   166- 185 (20) 1934
  219- 236 (18) 2453      |   212- 235 (24) 2530
  255- 276 (22) 2140 ++   |   252- 274 (23) 1386
  300- 319 (20)  915      |   299- 319 (21) 1299 ++
```

图1-25 用TMpred分析P51684序列所得到的7个可能跨膜螺旋区的相关性列表

3.) Suggested models for transmembrane topology

These suggestions are purely speculative and should be used with **extreme caution** since they are based on the assumption that all transmembrane helices have been found.
In most cases, the Correspondence Table shown above or the prediction plot that is also created should be used for the topology assignment of unknown proteins.

```
2 possible models considered, only significant TM-segments used

-----> STRONGLY prefered model: N-terminus outside
 7 strong transmembrane helices, total score : 14211
 # from   to length score orientation
 1   55   74 (20)    2707 o-i
 2   83  104 (22)    1914 i-o
 3  120  141 (22)    1451 o-i
 4  166  184 (19)    2170 i-o
 5  212  235 (24)    2530 o-i
 6  255  276 (22)    2140 i-o
 7  299  319 (21)    1299 o-i

------> alternative model
 7 strong transmembrane helices, total score : 12004
 # from   to length score orientation
 1   47   69 (23)    2494 i-o
 2   84  104 (21)    1470 o-i
 3  123  141 (19)    1352 i-o
 4  166  185 (20)    1934 o-i
 5  219  236 (18)    2453 i-o
 6  252  274 (23)    1386 o-i
 7  300  319 (20)     915 i-o
```

图1-26 用TMpred分析P51684序列所得到的7个可能跨膜螺旋区的跨膜拓扑模型

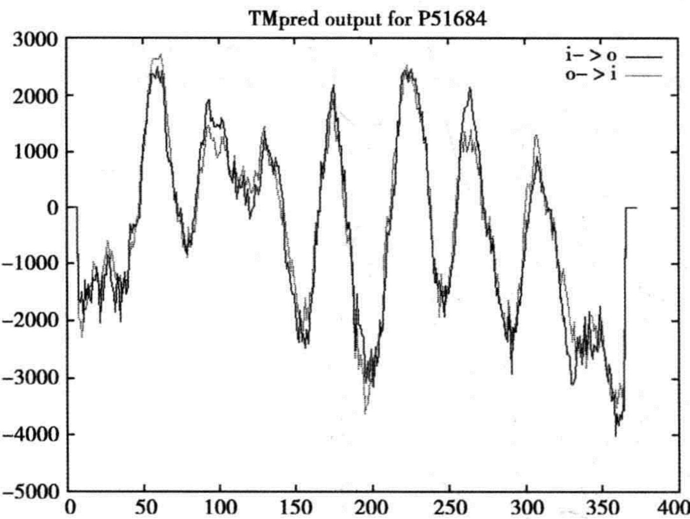


图1-27 用TMpred分析P51684序列所得到的7个可能跨膜螺旋区的图形显示结果

三、蛋白质信号肽的预测和识别

信号肽（signal peptide）是指新合成多肽链中用于指导蛋白质跨膜转移的末端（通常为N末端）氨基酸序列。信号肽在蛋白分泌的过程中起重要作用，分泌性蛋白质合成后由信号肽引导其穿过合成所在的细胞到其他组织细胞中。信号肽中至少含有一个带正电荷的氨基酸，中部有一个高度疏水区以通过细胞膜。信号肽假说认为，编码分泌蛋白的mRNA在翻译时首先合成的是N末端带有疏水氨基酸残基的信号肽，它被内质网膜上的受体识别并与之结合。信号肽经膜中蛋白质形成的孔道到达内质网内腔，并随机被位于腔表面的信号肽酶水解，由于它的引导，新生的多肽就能够通过内质网膜进入腔内，最终被分泌到胞外。信号肽的识别有助于蛋白质功能域的区分及蛋白质细胞定位。

前导肽（leader peptide）是信号肽的一种。在线粒体蛋白质的跨膜转运过程中，通过线粒体膜的蛋白质在转运之前大多数以前体形式存在，它由成熟蛋白质和N端延伸出的一段前导肽共同组成。迄今已有40多种线粒体蛋白质前导肽的一级结构被阐明，它们约含20~80个氨基酸残基，当前体蛋白跨膜时，前导肽被一种或两种多肽酶水解转变为成熟蛋白质，同时失去继续跨膜的能力。前导肽一般具有以下特性：①带正电荷的碱性氨基酸（特别是精氨酸）含量较为丰富，它们分散于不带电荷的氨基酸序列之间；②缺失带负电荷的酸性氨基酸；③羟基氨基酸（特别是丝氨酸）含量较高；④有形成两亲（既有亲水又有疏水部分）α-螺旋结构的能力。

可以利用因特网在线工具和信号序列捕获系统来判定基因序列中是否含有信号肽序列。SignalP是丹麦科技大学生物序列分析中心开发的信号肽及其剪切位点检测在线工具，其网址为http://genome.cbs.dtu.dk/services/SignalP/。该软件基于神经网络方法，用已知信号序列的革兰阴性原核生物、革兰阳性原核生物及真核生物的序列作为训练集。SignalP预测的是分泌型信号肽，而不是参与细胞内信号传递的蛋白质。

四、蛋白质的卷曲螺旋预测

卷曲螺旋是通过其疏水性界面相互缠绕在一起形成的一个十分稳定的结构，是控制蛋白质寡聚化的元件，它存在与很多蛋白质中，例如转录因子、病毒融合蛋白多肽等，在中间纤维中也有很长的这样的元件。

卷曲螺旋（coiled-coil）是存在于多种天然蛋白质中的一类由两股或者两股以上α螺旋相互缠绕而形成的平行或反平行左手超螺旋结构的总称。卷曲螺旋区域一般以7个氨基酸残基为单位组成，以a、b、c、d、e、f、g位置表示，其中a和d位置为疏水性氨基酸，其他位置的氨基酸残基为亲水性。卷曲螺旋是控制蛋白质寡聚化的元件，它存在与很多蛋白质中，许多含有卷曲螺旋结构的蛋白质具有重要的生物学功能，例如基因表达调控中的转录因子、病毒融合蛋白多肽等等，在中间纤维中也有很长的这样的元件。表1-24列出了常用的蛋白质卷曲螺旋预测工具。

COILS是由Swiss EMBNet维护的预测卷曲螺旋的在线工具，其网址为http://www.ch.embnet.org/software/COILS_form.html。该软件基于Lupas算法，在一个包含已知卷曲螺旋蛋

白结构的数据库中对查询序列进行搜索，同时也将查询序列与包含球状蛋白序列的PDB次级库进行比较，并根据两个库搜索得分决定查询序列形成卷曲螺旋的概率。COILS也可以下载到本地进行运算。

表1-24 常用的蛋白质卷曲螺旋预测工具

工具	网址	备注
Coiled-coil	http: //www.york.ac.uk/biology/units/coils/coilcoil.html	Mac
COILS	http: //www.ch.embnet.org/software/COILS_form.html	Web
EpitopeInfo	http: //epitope-informatics.com/Links.htm	Web

五、糖基化位点的预测与识别

糖基化是真核细胞中最常见的翻译后蛋白质修饰过程之一，在生物学过程中扮演着重要的角色，它能参与免疫防御、病毒复制、细胞生长等过程。蛋白质的糖基化有N-糖基化、O-糖基化、C-甘露糖糖基化以及糖基脂酰肌醇（GPI）锚区四种类型。其中O-糖基化参与很多细胞生化过程，诸如细胞黏附、细胞免疫、精卵结合、血液凝固以及微生物对细胞的黏附等。O-糖基化可调节细胞表面受体的表达和功能，从而影响生物细胞的生长和凋亡、胚胎发生等重要生命过程。但是O-糖基化位点的确切序列片段还不清楚，还未发现固定的模式，但是许多基于实验和计算的方法已经被应用在寻找糖基化和序列间的一致性。

NetOGlyc是由丹麦技术大学的生物序列分析中心维护的预测糖基化位点的在线工具，其网址为http: //www.cbs.dtu.dk/services/NetOGlyc/。NetOGlyc预测哺乳动物蛋白质中的糖基化位点，通过神经网络系统对序列进行分析，最后得到一个阈值分布和相应位点的得分，可以批量提交，也可以提交fasta格式的序列或者序列文件。

六、磷酸化位点的预测与识别

磷酸化是蛋白质重要的翻译后修饰之一，也是细胞调控的重要形式之一，磷酸化会影响到很多的细胞信号通路，包括代谢、生长、分化和膜运输等。由于磷酸化的重要性，磷酸化位点的理论识别成为计算生物学的重要研究内容。磷酸化位点附近存在保守残基片段，而这种保守性又与激酶类型相关。表1-25列出了常用磷酸化位点的预测与识别工具。

表1-25 常用的磷酸化位点的预测与识别工具

工具	网址	备注
KinasePhos	http: //kinasephos.mbc.nctu.edu.tw/	Web
GPS	http: //gps.biocuckoo.org/	Windows、Linux、Unix、Mac OS
pkaPS	http: //mendel.imp.ac.at/sat/pkaPS/	Web
NetPhos	http: //www.cbs.dtu.dk/services/NetPhos/	Web

第六节

应用实例

Section 6 Examples

真核生物的基因一般为断裂基因(interrupted gene),由内含子和外显子组成,编码序列通常被内含子隔开。虽然基因的结构式断裂的,但是mRNA的结构去不是断裂的。基因的初始转录物与基因的结构相同,经过mRNA剪接从mRNA前体中去除内含子得到信使mRNA。而对同一个mRNA前体,通过不同的剪接方式产生了不同的mRNA选择性剪接变体,使一个基因在不同时间、不同环境中能制造出不同的蛋白质,RNA的选择性剪切时高等真核生物基因中普遍存在的一种生命现象,它在真核基因表达调控中起着十分重要的作用。本节我们用实例介绍如何让利用NCBI的Spidey工具分析mRNA或者cDNA的外显子组成以及基因的选择性剪切分析。

一、利用Spidey工具识别mRNA/cDNA的外显子组成

图1-28是Spidey工具的序列在线提交页面,在主界面中有两个窗口,上方窗口用于输入基因组序列(直接粘贴序列或者用GenBank号); 下方窗口用于输入mRNA/cDNA序列(直接粘贴序列或者用GenBank号),可同时输入多条mRNA/cDNA序列与同一条基因组序列进行分析。"divergent sequences" 参数用于判断分析的序列间的差异;"Use large intron sizes" 参数表示是否接受默认的内含子长度限制,默认的内部内含子为35kb,末端内含子为100kb;"Genomic sequence" 参数用于判断序列的物种;"Out options" 参数用于选择结果输出的格式。

人类的FXYD5是一个重要的铁离子转运调节体,为了了解FXYD5的mRNA的外显子组成,我们在NCBI的GenBank数据库中检索到FXYD5的一条mRNA的记录号NM_014164,以及该基因所对应的基因组片段记录号AC002390,我们将AC002390填在上方的输入界面,NM_014164填入下方的输入界面,"Genomic sequence is" 参数选择 "Vertebrate",其余参数选择默认,然后点击 "Align" 开始分析,最后的结果以图形化的方式返回(图1-29)。结果显示FXYD5基因记录号为NM_014164的mRNA由9个外显子组成,结果详细给出了mRNA的每个外显子在基因组中对应位置和长度。同时Spidey工具在图形化结果的下方还给出了具体的序列比对的信息(图1-30)。

Genomic sequence (FASTA or GI/Accession):

Upload file: 选择文件 未选择文件

From: 0 To: 0

mRNA sequence(s) (One or more FASTA or GI/Accession):

Upload file: 选择文件 未选择文件

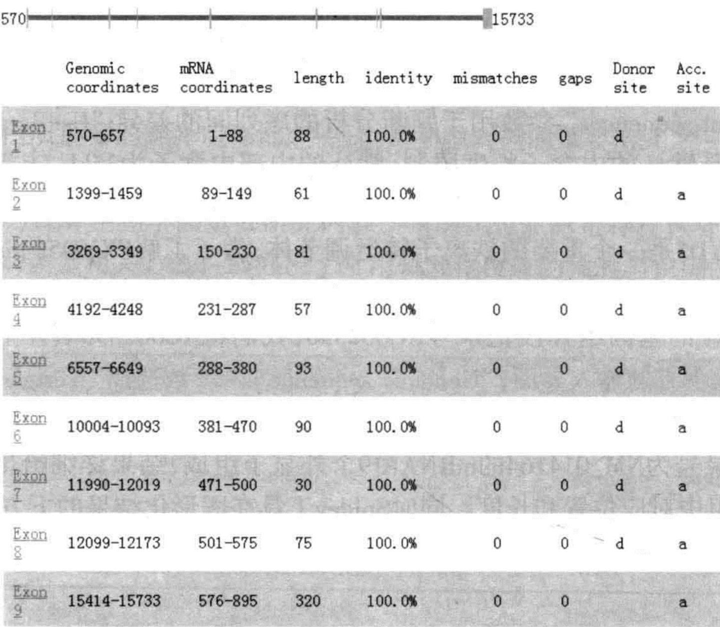
Align

Clear

☐ divergent sequences
☐ Use large intron sizes

Minimum mRNA-genomic identity 0 %
Minimum length of mRNA covered 0 %

图1-28 Spidey工具的序列在线提交页面



	Genomic coordinates	mRNA coordinates	length	identity	mismatches	gaps	Donor site	Acc. site
Exon 1	570-657	1-88	88	100.0%	0	0	d	
Exon 2	1399-1459	89-149	61	100.0%	0	0	d	a
Exon 3	3269-3349	150-230	81	100.0%	0	0	d	a
Exon 4	4192-4248	231-287	57	100.0%	0	0	d	a
Exon 5	6557-6649	288-380	93	100.0%	0	0	d	a
Exon 6	10004-10093	381-470	90	100.0%	0	0	d	a
Exon 7	11990-12019	471-500	30	100.0%	0	0	d	a
Exon 8	12099-12173	501-575	75	100.0%	0	0	d	a
Exon 9	15414-15733	576-895	320	100.0%	0	0		a

图1-29 Spidey工具的图形化结果显示

Exon 1: 570-657 (genomic); 1-88 (mRNA)

```
570    GCTCGCGCTCCCCTGGCCACACCCTCCGCCTGGACGCAGCAGCCACCGCC
                   ||||||||||||||||||||||||||||||||||||||||
1                  CCCTGGCCACACCCTCCGCCTGGACGCAGCAGCCACCGCC

610    GCGTCCCTCTCTCCACGAGGCTGCCGGCTTAGGACCCCCAGCTCCGACGT
       ||||||||||||||||||||||||||||||||||||||||||||||||
41     GCGTCCCTCTCTCCACGAGGCTGCCGGCTTAGGACCCCCAGCTCCGAC

       AAGTCCCT
```

Top

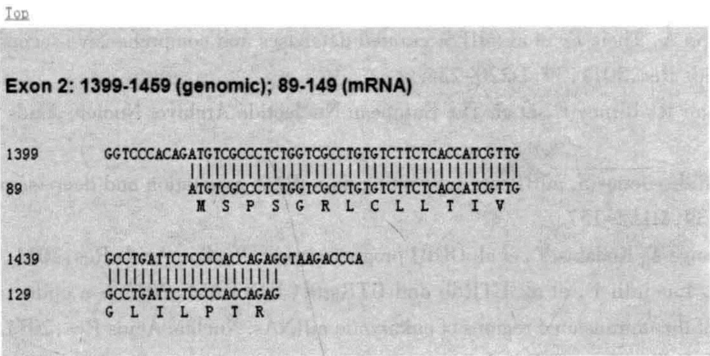


图1-30　Spidey工具的序列比对结果

二、利用Spidey工具进行可变性剪切的分析

NADPH氧化酶(nicotinamide vadenine dinucleotide phosphate oxidase，NOX)家族是许多非吞噬细胞中活性氧的主要来源，通过该途径产生的活性氧作为信号分子参与了细胞分化、增殖、凋亡等的调节。NOX1是NOX家族的一个成员，在GenBank数据库中检索发现了AF127763.2、AF166326.1、AF166327.1和AF166328.1四条非常相似的mRNA序列，这些序列可能是NOX1基因的可变性剪切产生的产物，将NOX1基因所在的基因组片段编号NG_012567.1黏贴在界面的上方输入框，四个mRNA的记录号粘贴在界面的下方输入框，参数选择默认，点击“Align”，最终显示的可变性剪切的图形界面结果(图1-31)。通过分析这个图形化结果和后续的序列比对的详细信息，我们将能很好地了解NOX1的可变性剪切的方式。

```
mRNA 1: gi|6138993|gb|AF127763.2| Homo sapiens mitogenic oxidase mRNA, complete cds
mRNA 2: gi|6138993|gb|AF127763.2| Homo sapiens mitogenic oxidase mRNA, complete cds
mRNA 3: gi|6672077|gb|AF166327.1| Homo sapiens NADPH oxidase homolog 1 long form
(NOH1) mRNA, alternatively spliced, complete cds
mRNA 4: gi|6672079|gb|AF166328.1| Homo sapiens NADPH oxidase homolog 1 long form
variant (NOH1) mRNA, alternatively spliced, complete cds
```

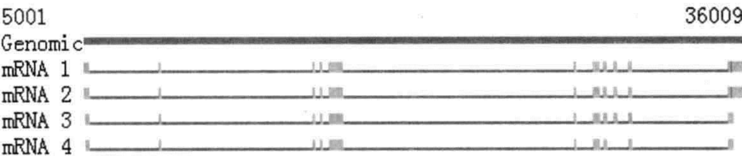


图1-31　Spidey工具进行可变性剪切分析的图形化结果

(周　猛)

参考文献

1. Sayers EW, Barrett T, Benson DA, et al. Database Resources of the National Center for Biotechnology Information. Nucleic Acids Res, 2011, 39, D38-51.
2. Petersen TN, Brunak S, von Heijne G, et al. SignalP 4. 0: discriminating signal peptides from transmembrane regions. Nat Methods, 2011, 8, 785-786.
3. Mewes HW, Ruepp A, Theis F, et al. MIPS: curated databases and comprehensive secondary data resources in 2010. Nucleic Acids Res, 2011, 39, D220-224.
4. Leinonen R, Akhtar R, Birney E, et al. The European Nucleotide Archive. Nucleic Acids Res, 2011, 39, D28-31.
5. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res, 2011, 39, D152-157.
6. Kaminuma E, Kosuge T, Kodama Y, et al. DDBJ progress report. Nucleic Acids Res, 2011, 39, D22-27.
7. Grillo G, Turi A, Licciulli F, et al. UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res, 2011, 38, D75-80.
8. Gardner PP, Daub J, Tate J, et al. Rfam: Wikipedia, clans and the "decimal" release. Nucleic Acids Res, 2011, 39, D141-145.
9. Benson DA, Karsch-Mizrachi I, Lipman DJ, et al. GenBank. Nucleic Acids Res, 2011, 38, D46-51.
10. Pruitt KD, Tatusova T, Klimke W, et al. NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Res, 2009, 37, D32-36.
11. Chan PP, Lowe TM. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res, 2009, 37, D93-97.
12. Szymanski M, Erdmann VA, Barciszewski J. Noncoding RNAs database (ncRNAdb). Nucleic Acids Res, 2007, 35, D162-164.
13. Karro JE, Yan Y, Zheng D, et al. Pseudogene. org: a comprehensive database and comparison platform for pseudogene annotation. Nucleic Acids Res, 2007, 35, D55-60.
14. Gelfand Y, Rodriguez A, Benson G. TRDB—the Tandem Repeats Database. Nucleic Acids Res, 2007, 35, D80-87.
15. Wu CH, Apweiler R, Bairoch A, et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res, 2006, 34, D187-191.
16. Wu CH, Nikolskaya A, Huang H, et al. PIRSF: family classification system at the Protein Information Resource. Nucleic Acids Res, 2004, 32, D112-114.
17. Ursing BM, van Enckevort FH, Leunissen JA, et al. EXProt: a database for proteins with an experimentally verified function. Nucleic Acids Res, 2002, 30, 50-51.
18. Ruitberg CM, Reeder DJ, Butler JM. STRBase: a short tandem repeat DNA database for the human identity testing community. Nucleic Acids Res, 2001, 29, 320-322.
19. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res, 1994, 22, 4673-4680.
20. Lupas A, Van Dyke M, Stock J. Predicting coiled coils from protein sequences. Science, 1991, 252, 1162-1164.
21. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol, 1990, 215, 403-410.
22. H. D. Huang, T. Y. Lee, S. W. Tseng, et al. "KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites" Nucleic Acids Research, Vol. 2005, 33, W226-229.
23. Hofmann K, Stoffel W. TMbase-A database of membrane spanning proteins segments. Biol. Chem. Hoppe-

Seyler, 1993, 374(47): 166.
24. 吴宪明,吴松峰,任大明,等. 密码子偏性的分析方法及相关研究进展. 遗传. 2007, 29(4): 420–426.
25. 李衍达,孙之荣. 生物信息学基因和蛋白分析的实用指南. 北京: 清华大学出版社; 2000年.
26. 胡松年. 基因表达序列标签(EST)数据分析手册. 杭州: 浙江大学出版社; 2005年.
27. 薛庆中. DNA和蛋白质序列分析工具. 北京: 科学出版社; 2009年.
28. 孙啸,陆祖宏,谢建明. 生物信息学基础. 北京: 清华大学出版社; 2005年.
29. 李霞,李亦学,廖飞. 生物信息学. 北京: 人民卫生出版社; 2010年.

第二章

CHAPTER 2

基因芯片数据分析

MICROARRAY DATA ANALYSIS

基因芯片(gene chip)通常被称为微阵列(microarray),它是20世纪90年代发展起来的一种高通量检测基因表达水平的生物技术。基因芯片数据能够反映生物个体的所有基因在特定组织、器官、生理状态(如,疾病)或发育阶段中的表达情况。基因芯片技术已经被广泛应用到基因的功能研究、基因的转录调控分析、疾病标志物(marker)的识别、疾病亚型的确定、疾病的精确分类以及药物靶点的筛选等领域,为复杂疾病的分子机制研究提供了转录水平的全局性视角,极大加快了药物研发的进程以及个性化医疗的开展。本章将首先简要介绍各种常见的基因芯片平台;然后结合具体软件的使用,讲解基因芯片数据的预处理方法和主要分析技术(如:特征基因的识别、聚类和分类分析等);最后通过两个具体实例说明基因芯片数据在人类复杂疾病研究中的应用。

第一节

基因芯片平台简介

Section 1 Introduction to Microarray Platforms

基因芯片是由大量cDNA或寡核苷酸探针密集排列而形成的探针阵列，其基本原理是杂交测序方法。基因芯片的主要特点是微型化（芯片小巧）、集约化（一张芯片可以完成大量基因的检测）、自动化（一次动作可以完成实验室从探针的固定、探针与样本杂交等过程多个步骤的工作）、平行化（检测基因在同一时空状态下的表达）、快速灵敏（检测时间一般可在30分钟内完成，如果采用控制电场的方式，杂交时间可控制在1分钟左右）、样品用量少、成本相对低廉等优点。基因芯片的类型众多，大致有以下几种分类方式：①以基质材料分，有尼龙膜、玻璃片、硅胶晶片、微型磁珠等；②以所检测的生物信号种类分，有核酸、蛋白质、生物组织碎片甚至完整的活细胞；③按工作原理分类，有杂交型、合成型、连接型、亲和识别型等。以下将介绍几种具有代表性的基因芯片的制备过程及特点。

一、cDNA芯片

cDNA芯片（cDNA microarray）是在1995年由美国stanford大学首先研制成功的。cDNA芯片的制作流程如图2-1所示：首先通过克隆的方法获得目标cDNA序列，将其作为探针高密度固定在基质上制备cDNA芯片（探针的序列是已知的）；然后从待检测的实验细胞（experimental cell）和对照细胞（control cell）中分别提取总mRNA，由于RNA本身不稳定，而cDNA保存时间较长，因此，将mRNA反转录（reverse transcription）成cDNA，并分别用红色荧光分子（Cy5）和绿色荧光分子（Cy3）进行标记；接下来将两组cDNA样本等比例混合，在一定的实验条件下与芯片上的探针进行杂交，杂交完成后洗脱没有与探针互补结合的cDNA片段；最后，将芯片置于黑箱中，对芯片进行激光共聚焦扫描，获得每个探针杂交后的荧光强度。如果基因在两组细胞中的表达水平相同，其扫描后的图像为黄色，如果基因在两组细胞中的表达水平不同，则扫描后的图像呈现红色或绿色，荧光强度值定量反映了基因的相对表达水平。此外，由于cDNA芯片的探针来源于样本的cDNA克隆，因此探针的长短不一，需要的杂交条件也不同，但在进行芯片杂交时只能设定一个杂交条件，其结果可能会出现由于实验本身导致的非特异性杂交和杂交效能低等问题，因此，cDNA芯片的可靠性和重复性不是很理想。

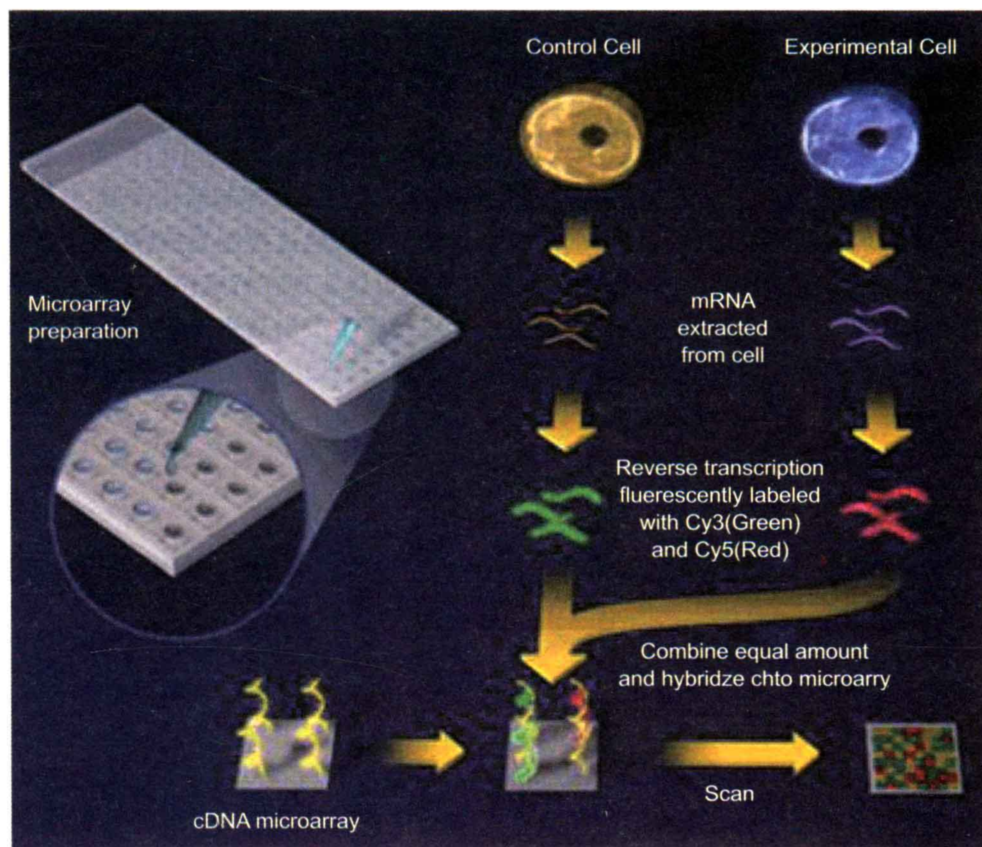


图2-1　cDNA芯片制作流程

来源于: http: //www.scq.ubc.ca/spot-your-genes-an-overview-of-the-microarray/

二、寡核苷酸芯片

寡核苷酸芯片(oligonucleotide microarray)的主要原理与cDNA芯片类似，主要通过碱基互补配对的原则进行杂交，来检测探针所对应基因的表达水平。寡核苷酸芯片的探针并不是来源于cDNA克隆，而是预先设计并合成的、能够代表每个基因的特异性寡核苷酸片段，长度约为50bp，然后将其点样到特定的基质上制备成芯片，从而克服了cDNA探针序列太长导致的非特异性交叉杂交和探针杂交条件不同所导致的数据结果不可信。然而，由于寡核苷酸探针是预先一次性合成，并且每次芯片制备中探针的消耗量又很少，而在一系列实验中芯片的制备存在时间差，因此早期合成的寡核苷酸片段可能存在降解的情况，从而导致最终检测质量的下降。

三、原位合成芯片

原位合成芯片(light-controlled in situ synthesis of DNA microarrays)采用的是光引导聚合技术，在芯片的特定部位原位合成寡核苷酸探针，其方法不同于上述两种芯片的点样技术。

原位合成芯片的探针制备过程如图2-2所示：先将基片支持物（wafer）羟基化，然后用对光敏感的保护基团将羟基基团保护起来。每次选取特制的光掩膜（mask）覆盖在基片上，遮挡不需要合成的部位。当光通过光掩膜照射到基片上时，需要聚合的部位透光，受光照射部位的羟基去保护而活化。然后加入3' 端活化（5' 羟基末端连接光敏保护基团）的单一核苷酸单体底物后，发生偶联反应。在一轮反应之后更换另一张光掩膜来控制活化区域，并换另一种核苷酸单体实现在特定位点合成预定的序列寡聚体。每次通过控制光掩膜（决定哪些区域应被活化）以及所用核苷酸单体的种类和反应次序就可以实现在特定位点合成大量预定序列寡聚体的目的。使用多种掩盖物能以更少的合成步骤生产出高密度的阵列，在合成循环中探针数目呈指数增长。光掩膜的设计和严格的工艺流程使制造的芯片具有高密度、高重复性和一致性。

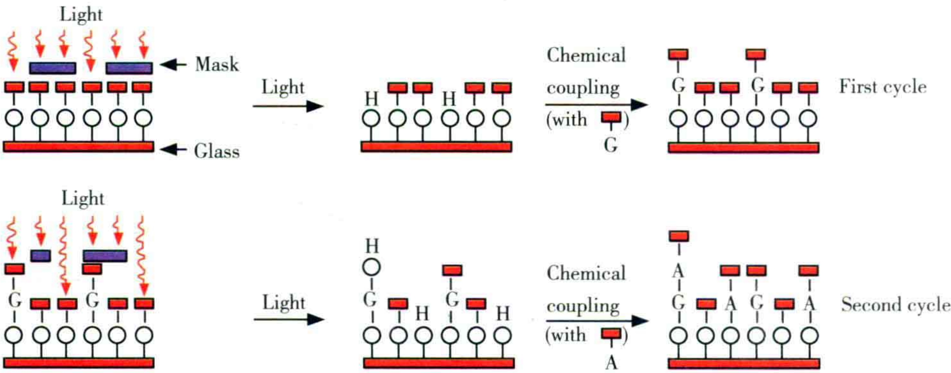


图2-2 原位合成芯片的探针制备过程

来源于：http://bioservices.capitalbio.com/fwpt/Affymetrixpt/3943.shtml

原位合成芯片为单通道芯片，即只使用一种荧光分子对样本进行标记，通过检测荧光强度获得基因的表达水平。由于寡核苷酸探针长度较短（一般为15~25个碱基），对于某个待检测的基因通常需要设计多个相互重叠的探针构成探针集，从而有效减少探针杂交非专一性的影响。该类芯片的主要优势在于所有探针都是在一个条件下完成的，因此同一批芯片的探针浓度均一性较好；此外，由于该类芯片的探针合成和芯片制备是同时进行的，探针不需要预先合成，所以避免了点样芯片中探针的降解情况，从而保证了实验的重复性；同时，考虑到探针的非特异性杂交问题，该类芯片通常针对每段参考序列（reference sequence）设计一对寡核苷酸探针，其中一个是完全匹配（perfect match，PM）的探针，另一个是中间有一个碱基错配（mismatch，MM）的探针，计算时将每对PM-MM探针的检测信号综合起来，这样有助于区分特异性结合与非特异性结合的靶片段，从而提高探针灵敏度和特异性（图2-3）。这种PM-MM设计对于复杂序列背景样品中低丰度表达产物的检测有明显优势。

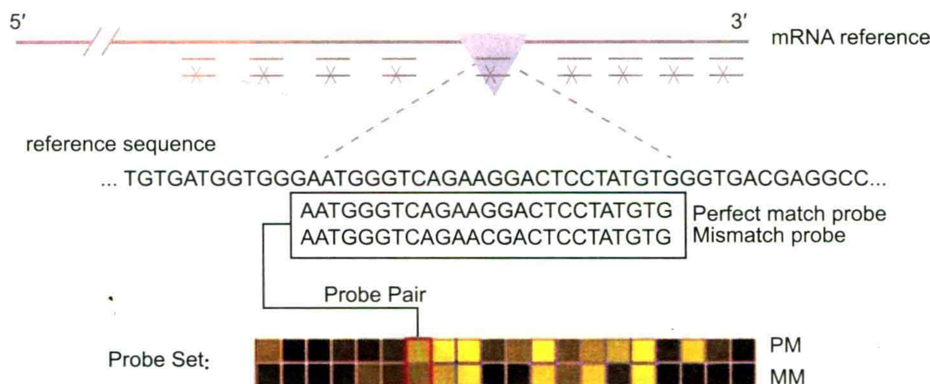


图2-3 原位合成芯片的探针设计原则

来源于：http: //bioservices.capitalbio.com/fwpt/Affymetrixpt/3943.shtml

四、光纤微珠芯片

光纤微珠芯片（beadarray microarray）是利用无孔无荧光硅珠阵列（beadarray）的技术制备的一种芯片，是新一代的基因芯片。光纤微珠芯片具备如下特点：①密度高，该类芯片是目前最高密度的芯片制备技术，每平方厘米有约400万个点（微珠）；②上样量低，每个芯片在一轮反转录的情况下仅需50~100ng RNA；③重复性高，芯片设计中的“无序自组装”方式以及每种类型微珠的30倍重复的特点保证了芯片的高重复性；④数据准确性高，定量PCR（qPCR）是检验芯片数据的黄金标准，实验表明全基因组表达芯片（Human-6，Humanref-8）和qPCR相关系数R^2=0.93，特定基因组研究芯片和qPCR相关系数R^2=0.97；⑤100%质量控制，芯片生产过程中采用专利的解码技术，质量控制能深入到每张芯片上每个微珠的每个特性，保证数据的可靠性和重复性；⑥性价比高，光纤微珠芯片价格是传统商品化芯片的1/2~1/3，有效降低了芯片成本。

光纤微珠芯片设计的基本原理如图2-4所示，其主要组成元件是光导纤维和纳米材料（微珠），探针连接在微珠上，每个探针由两部分组成：23bp的地址序列（address）和50bp的探针序列。地址序列对每种微珠进行编码，特异对应于某个微珠，而探针序列则代表某个基因的特异片段。探针在合成纯化后与微珠通过化学反应连接，每个微珠可以连接100万左右相同的探针，为保证充足的探针数目，每个微珠还设计了30个重复。将不同类型的微珠进行混合形成“微珠池”，将若干束微小光纤插入微珠池，每5万根光纤组成一束，每根光纤的末端有一个用化学方法蚀刻的微孔，每个微孔内恰好仅可容纳一个直径为3μm的微珠，微珠以“无序自组装”的方式随机进入光纤束上的微孔组装成芯片。将从样本中提取的mRNA反转录成cDNA，并进一步产生cRNA（通过掺入带荧光标记的核苷酸进行标记），带有标记的cRNA与微珠上的特异性探针杂交。从激光扫描仪上发出激光通过光纤传递给荧光素，后者发出的光又通过光纤传递给检测器。最后，采用解码流程对芯片上微珠的类型、位置、数量、信号强弱进行解读，如果某个微珠的质量控制不达标，该通道将被关闭。解码过程同时完成了对芯片信息的采集以及100%的质量控制。

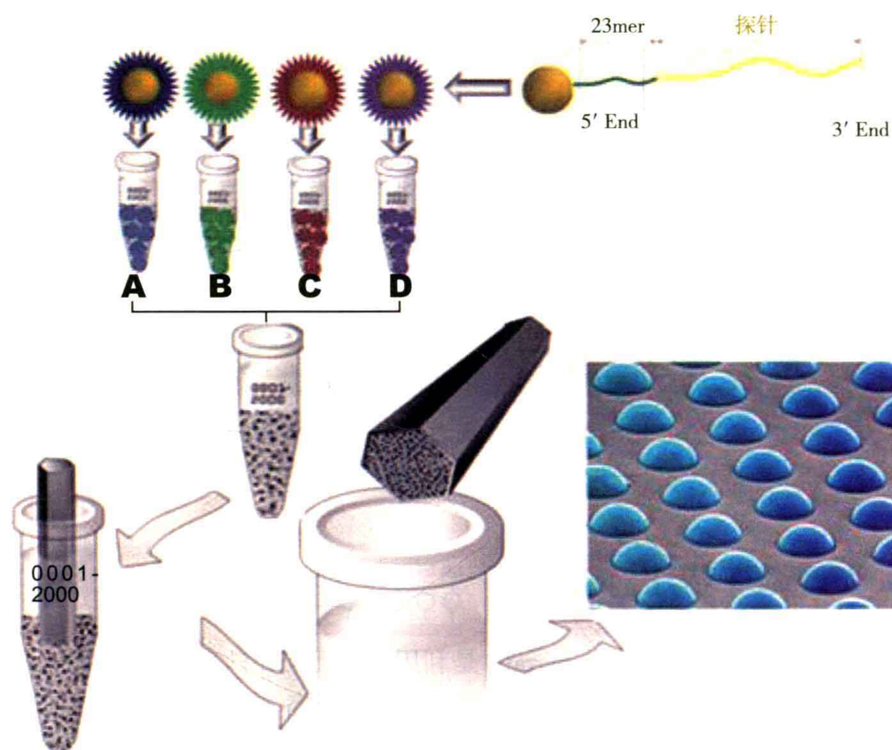


图2-4 光纤微珠芯片设计原理

第二节 基因芯片数据的预处理

Section 2　Preprocessing of Microarray Data

由于基因芯片平台的差异、系统误差的存在以及后期计算的需要，在对基因芯片数据进行聚类、分类等分析之前，往往需要先进行预处理(pre-procession)。预处理的过程主要包括数据提取，将高通量的荧光信号转化成基因表达数据；数据过滤，去除异常数据和噪声数据；补缺失值，保证数据的完整性；对数转换，以满足正态分布的分析要求；标准化，纠正系统误差，以发现真正的生物学变异。

一、基因芯片数据的提取

双通道芯片使用Cy5(红)和Cy3(绿)两种荧光分别标记实验样本和对照样本的cDNA序列，然后杂交至同一芯片。用不同波长的激光扫描芯片，获得荧光强度值。每个荧光点的原始信号值包括前景值和背景值，该点的荧光强度则用前景值减去背景值表示。cDNA芯片扫描得到的结果反映了基因在实验样本和对照样本中的相对表达水平。在扫描后的芯片图像上，红色表示该点所检测的基因在实验样本中表达呈现上调，绿色表示表达下调，黄色表示表达无改变。对于单通道芯片，扫描后的荧光强度由深到浅依次为蓝黑、蓝、高蓝、绿、黄、橙、红、白。颜色越深表示荧光强度越高，即与探针杂交的RNA越多，从而基因的表达量越高。芯片扫描系统的图像处理软件包括将荧光信号转化成数字信号的数据提取过程和基于探针集的基因表达值汇总过程。提取后的基因表达数据可以用矩阵形式表示，行代表基因，列代表样本，矩阵中的元素表示基因在样本中的表达水平，这种类型的数据通常被称为基因表达谱(gene expression profile)。

二、数据过滤

数据过滤是数据分析前必须进行的一项工作。基因芯片中每个点的信号强度是前景信号值减去背景信号值，然而，有时会出现负值或很小的值，显然负值是没有生物学意义的。另外，由于过闪耀现象、物理因素导致的信号污染、杂交效能低或点样问题等因素都可能导致数据的不真实。数据过滤的目的就是要去除表达水平是负值、很小的数据或者明显的噪声数据，通常的处理方法是将它们置为缺失、赋予统一的数值或者去除。

三、补缺失值

基因表达水平过高或过低都会影响荧光强度的检测，导致数据缺失；其他因素，例如芯片图像的损坏、指纹、灰尘等原因，也会产生缺失值。数据的缺失对于后续的数据分析有着很大的影响，为了保证基因表达数据的完整性，补缺失值是十分必要的，常用的方法有直接删除、补均值、k近邻法和回归法等。

最简单的方法就是直接删除含有缺失值的行向量或列向量；或者计算每行或每列中含有的缺失值数目，如果缺失值过多，则删除此行或列，否则用0、每行或每列的均值或中值进行补缺。但用此方法补出的缺失值很难评估其与真实值的接近程度，因此，还可以用k近邻法和回归法等来估算缺失值。

k近邻法的基本思想是利用与待补缺基因距离最近的k个邻居基因的表达值推测待补缺基因的表达值。首先确定含有缺失值的基因的k个邻居，然后利用邻居基因在该样本中的加权平均估计缺失值，常用的定义邻居基因的距离函数有欧式距离或相关系数。

回归法与k近邻法相似，首先确定待补缺基因的k个邻居，然后利用每个邻居基因分别作线性回归模型预测缺失值，最后将k个缺失值加权求平均作为最终缺失值的估计值。基本步骤为：

1. 确定含有缺失值的基因i的k个邻居基因，设$X_1,X_2,\cdots,X_k$为基因i的k个邻居在n个样本中的表达水平。

2. 具有缺失值的基因i较之邻居基因分别作线性回归模型：

$$\begin{aligned} X_i &= a_1 + b_1 X_1 \\ X_i &= a_2 + b_2 X_2 \\ &\vdots \\ X_i &= a_k + b_k X_k \end{aligned} \quad (2\text{-}1)$$

3. 基于回归模型预测缺失值（假设基因i在样本j中表达值缺失）：

$$\begin{aligned} x_{i,j}^1 &= a_1 + b_1 x_{1,j} \\ x_{i,j}^2 &= a_2 + b_2 x_{2,j} \\ &\vdots \\ x_{i,j}^k &= a_k + b_k x_{k,j} \end{aligned} \quad (2\text{-}2)$$

4. k个缺失值的加权平均作为最终缺失值的估计值：

$$x_{i,j} = \sum_{g=1}^{k} w_g x_{i,j}^g \quad (2\text{-}3)$$

这里w_g为邻居基因的权重，若邻居基因与基因i的距离近，则权重大，反之权重小。

四、数据对数化处理

基因芯片数据一般呈偏态分布，影响数据的进一步分析。将数据对数化转换后，数据

会近似服从正态分布(图2-5),从而为后续的数据分析带来方便,通常取以2为底的对数转换。

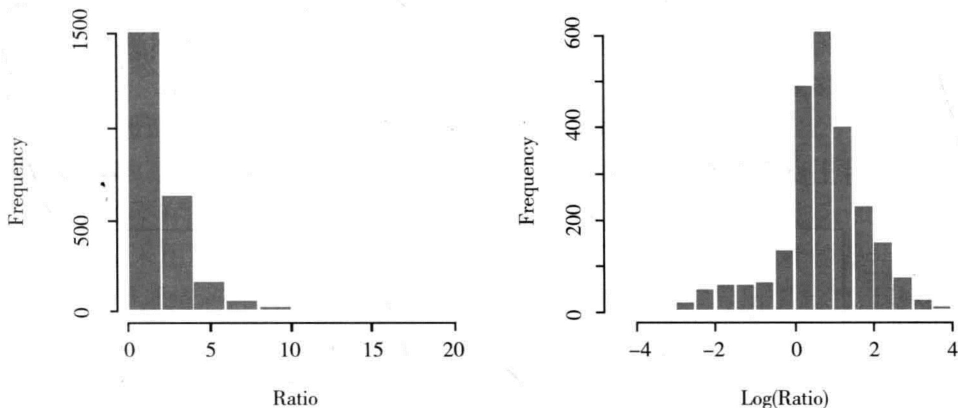


图2-5 数据对数转换前后log-Ratio值分布图

五、数据标准化

由于实验条件、基质、染料物理特性、点样针以及扫描仪采集数据时的参数设置等存在着差异,因此,基因芯片实验不可避免地产生了一些系统误差。数据标准化(normalization)的目的就是去除这些系统误差,从而挖掘出真正的生物学变异,确保后期数据分析的可靠性。

在对芯片进行标准化处理时,通常涉及参照基因的选择问题,那么哪些基因适合作为参照物呢?一般是以具有稳定表达的基因作为芯片标化的参照基因,这些基因在不同条件下表达值相同,因此,测得基因的荧光强度值的差异主要是由系统误差造成的,这样便可估计出系统误差的大小。稳定表达的基因主要有以下几种:持家基因和人工合成的控制基因可以作为参照基因,但是由于实验误差以及杂交特异性的问题,它们通常并不像人们想象的那样在不同实验条件下稳定表达,这就使得标准化结果的可靠性不高;此外,在基因芯片中,真正表达异常的基因只有一小部分,大部分基因在不同条件下都是稳定表达的,所以运用这大部分稳定表达的基因作为参照基因,标准化结果更为可靠。

由于单、双通道芯片制作原理不同,系统误差的来源也不同,所以在进行数据标准化时需要选用不同的方法。

(一) cDNA芯片

cDNA芯片的数据标准化主要分为片内标准化和片间标准化。片内标准化是对一个实验中的不同芯片进行独立操作,一般指去除荧光染色和点样针带来的系统误差。片间标准化的目的是去除不同芯片间的系统误差,使不同芯片检测的基因表达值具有可比性。

1. 片内标准化　片内标准化的主要方法有全局标准化、荧光强度依赖的标准化和点样针组内标准化。本节重点介绍全局标准化方法,主要过程如下:

cDNA芯片检测的荧光强度值表示的是基因的相对表达水平,取对数后(log-Ratio值)近

似服从正态分布。由于芯片上大部分基因的表达都是稳定的，所以芯片上所有基因的log-Ratio值均值应该为0（图2-6黄线所示）。而实际上，由于红光和绿光的荧光强度存在差异，即使表达完全相同的两个基因经Cy5和Cy3标记后所测得的荧光强度也不一致，因此，log-Ratio值分布的均值会偏离0（图2-6红线所示）。全局标准化的目的就是将实际测得的log-Ratio值分布的峰值位置移至0处，公式如下：

$$\log_2 R/G \rightarrow \log_2 R/G - c \quad (2\text{-}4)$$

其中，c表示芯片上所有基因的log-Ratio的中值或均值。

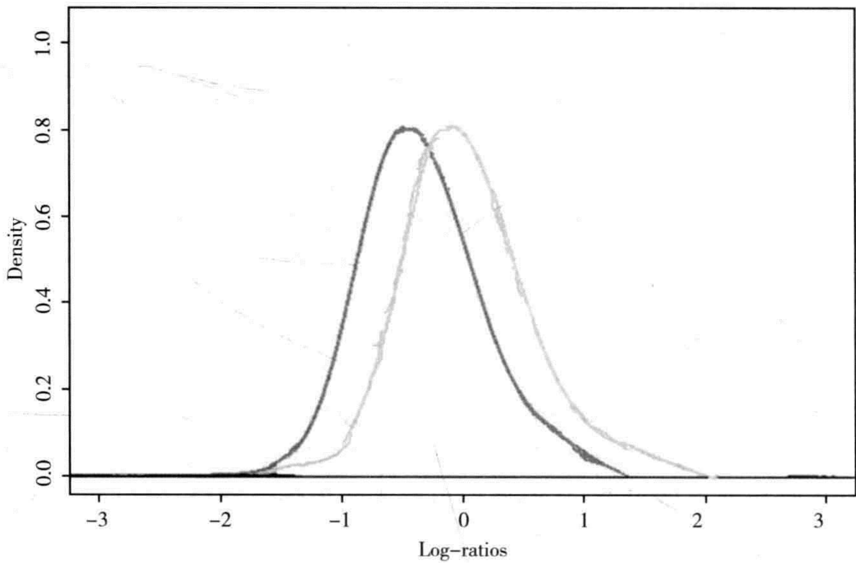


图2-6 全局标准化前后log-Ratio值分布图

来源于：Yang et al.: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.Nucleic Acids Res.,2002,30(4): e15.

全局标准化消除了染料偏倚带来的系统误差，应用较为普遍。在芯片实验中，染料偏倚的大小还依赖于荧光强度的高低，在不同的荧光强度下，对应的log-Ratio值分布的峰值偏离0的大小也不同。因此，荧光强度依赖的标准化的目的就是将不同荧光强度对应的log-Ratio值分布的峰值移回0处，消除荧光强度依赖的染料偏倚。此外，由于点样针的长短、粗细、磨损程度、点样顺序等差异的存在，也会引入系统误差，点样针组内标准化正是为了消除这种点样针带来的系统误差。

2. 片间标准化　片间标化的常用方法有分位数标准化和中位数标准化。

（1）分位数标准化：分位数标化的前提假设是每张芯片所检测的数据具有相同的分布，具体算法如下：

1）将基因表达谱中的每列（每张芯片）数据分别按照从大到小排序。

2）在排序后的矩阵中，每行每个位置的数据均用该行的均值所替代。

3）将新矩阵的每列数据分别按照在原矩阵中的位置重新排序，从而得到标准化的基因表达谱。

将芯片进行分位数标准化后就能保证每张芯片具有完全相同的数据分布。

（2）中位数标准化：对于双通道芯片数据来说，中位数标准化方法就是将每张芯片上的数值减去各自芯片上log-Ratio值的中位数。通过这样的处理，所有芯片的log-Ratio值的中位数就都变成了0，从而使得不同芯片间的log-Ratio值具有可比性。

（二）单通道芯片

由于单、双通道芯片的制备原理不同，系统误差的来源也不同。相比于双通道的cDNA芯片，与单通道的寡核苷酸芯片杂交的是单个样本，而不是实验样本与对照样本的混合物，所以单通道芯片不存在cDNA芯片中所涉及的染料偏倚所带来的系统误差；此外，单通道芯片的探针一般是采用原位合成的方法而非点样法，所以也不存在点样针的差异所产生的系统误差。因此，单通道芯片的系统误差主要是由不同芯片间的差异所引起的，其标准化方法与双通道的标准化方法类似的，这里不再单独介绍。

六、应用举例

BRB-Arraytools是基因芯片数据预处理的常用软件之一（详细的功能描述见本章第六节）。在此，我们以一套阿尔茨海默病相关的基因表达谱数据为例来详细讲解如何利用该软件进行数据预处理，该套数据是利用Affymetrix公司的寡核苷酸芯片HG-U133 Plus 2.0 Array检测阿尔茨海默病病人和正常老年人大脑中六个不同区域的基因表达情况，其在GEO（gene expression omnibus）数据库中的编号为GSE5281。我们仅选择其中一个区域——内侧颞回（middle temporal gyrus，MTG）的数据来进行说明，具体操作步骤如下：

第一步：导入芯片数据（图2-7）。使用“Import data”下的“General Format Importer”导入基因芯片数据文件，在该文件中数据之间应为Tab键分隔（或使用Excel文件），此外，也可使用“Data Import Wizard”进行导入。

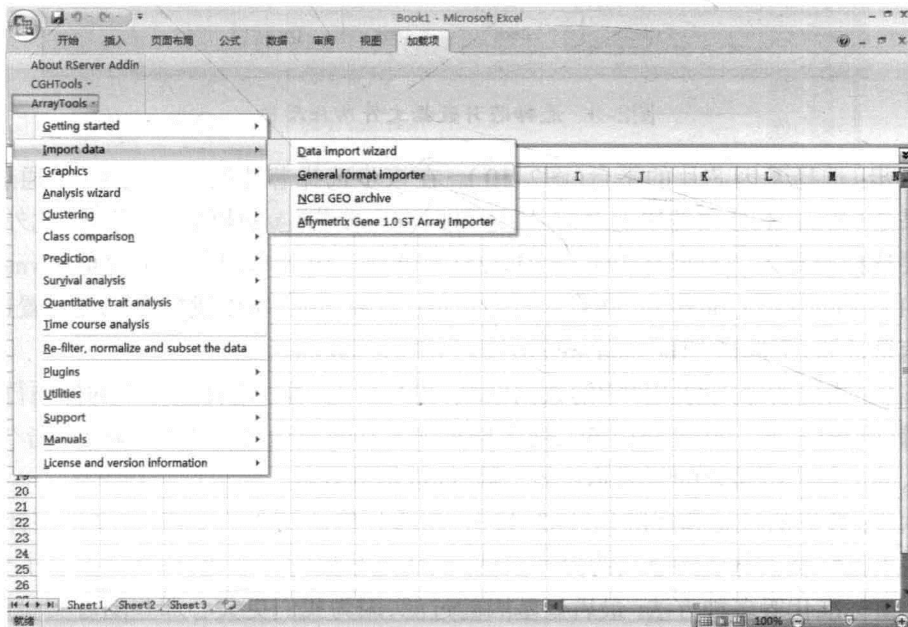


图2-7 导入芯片数据

第二步：选择文件类型（图2-8）。如果需要导入的基因芯片数据是每张芯片用单独的文件存储，多个文件保存在一个文件夹中，则选择"Arrays are saved in separate files stored in one folder"；如果多张芯片数据组织成一个矩阵的形式，存储在一个文件中，则选择"Arrays are saved in a horizontally aligned file"，由于本例的数据是存储在一个文件中的基因表达谱数据，因此，我们选择后者。

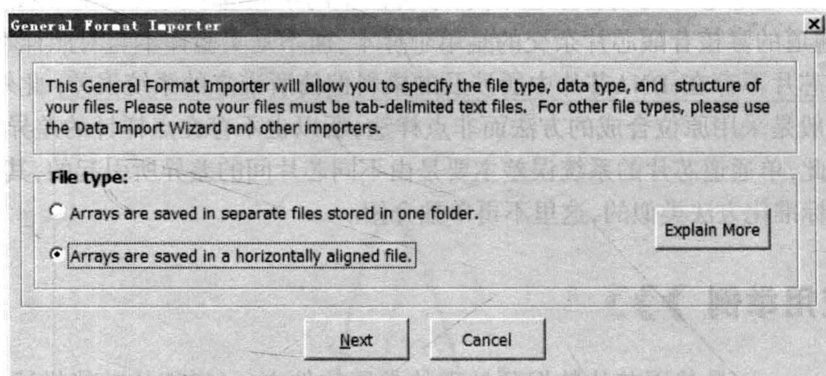


图2-8 选择文件类型

第三步：选择芯片数据文件所存储的路径（图2-9，注意路径中不能包含中文）。

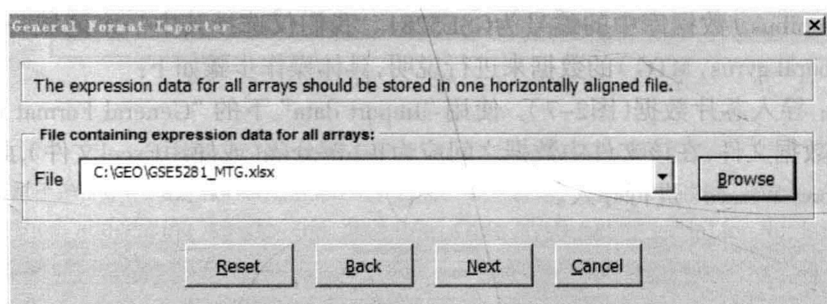


图2-9 选择芯片数据文件所在路径

第四步：选择基因芯片的平台（图2-10）。在该步选择芯片的平台类型（单通道或双通道），如果是Affymetrix公司的单通道芯片，还可以进一步指定具体的平台型号，此外，在该步骤还需要选择所导入的数据是否进行了log2对数转换。由于本例中采用的是Affymetrix公司的HG-U133 Plus 2.0 Array平台，并且未进行过log2对数转换，所以我们在相应位置选择具体的平台信息，同时不选择"The data are already log2 transformed."。

第五步：指定所导入的文件中的数据区域（图2-11）。通过选择文件中的标题行、第一行数据、探针所在列、第一列数据和第二列数据来确定基因表达谱的数据区域，点击"Next"会显示导入的文件中所包含的基因芯片的个数，即数据的列数。

第六步：数据的过滤和标准化（图2-12）。该部分包括三个子步骤，首先是探针的过滤，删除那些表达强度很低或无意义的探针数据；然后是数据的标准化，该软件整合了分位数标准化和中值标准化等多种方法；最后是基因的过滤，因为我们更关心那些随着实验条件的改变表达水平发生变化的基因，因此，在这步可以将那些表达波动较小的基因去除。

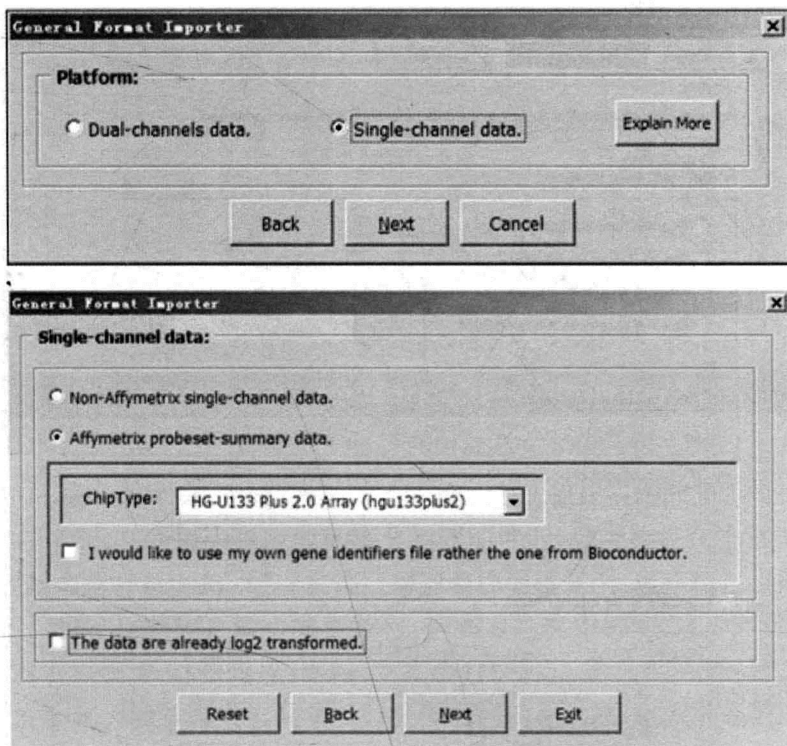


图2-10 选择基因芯片的平台

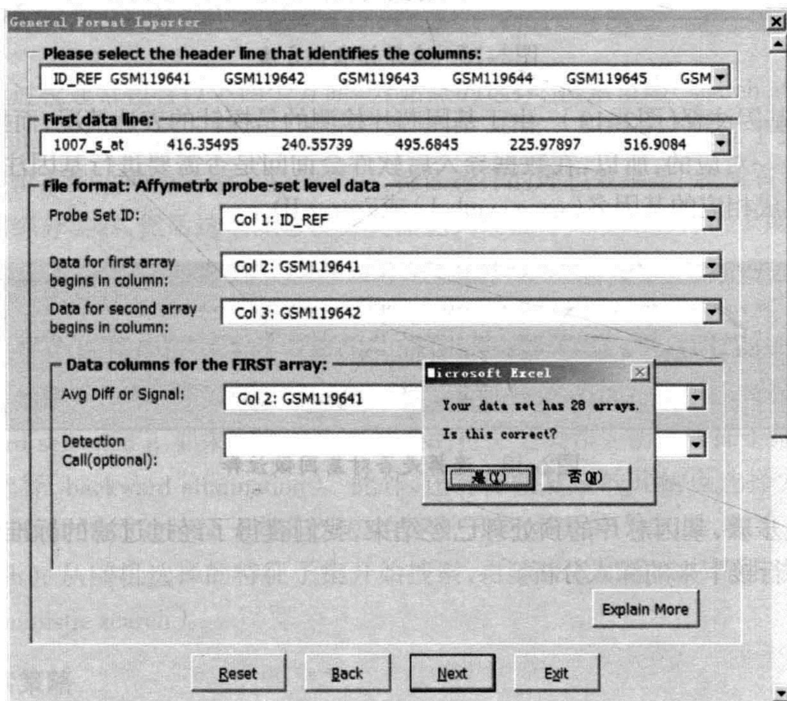


图2-11 选择文件格式

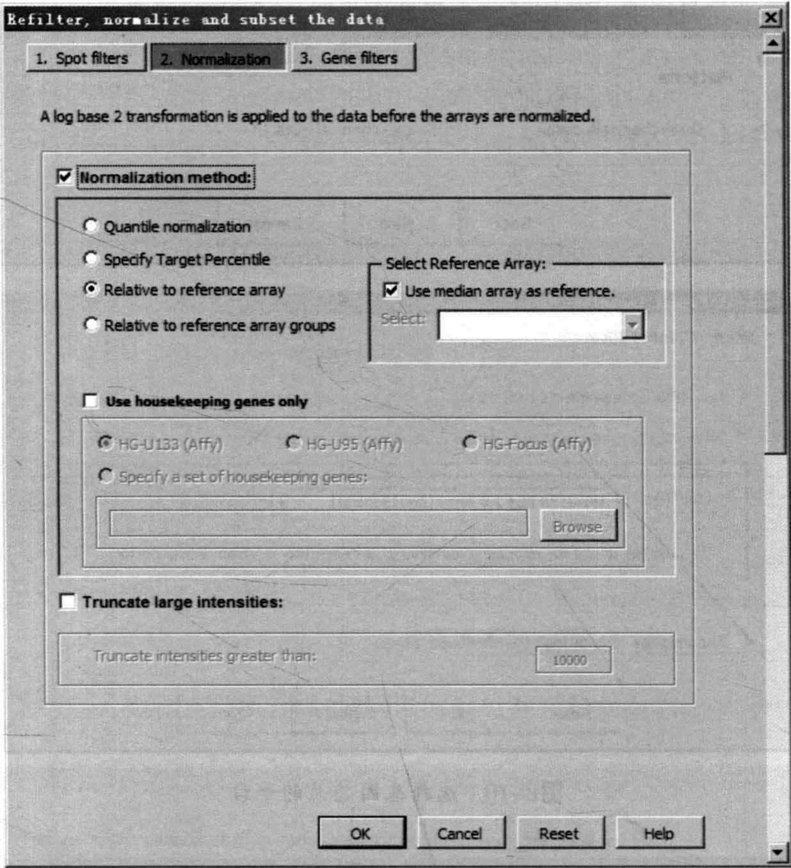


图2-12 选择标准化方法

第七步：基因注释(图2-13)。由于基因芯片检测的是探针的表达情况，而探针和基因之间往往不是一一对应的，所以，在数据导入后软件会询问是否需要进行基因注释，即是否需要将探针转换成相应的基因名(gene symbol)或Entrez ID。

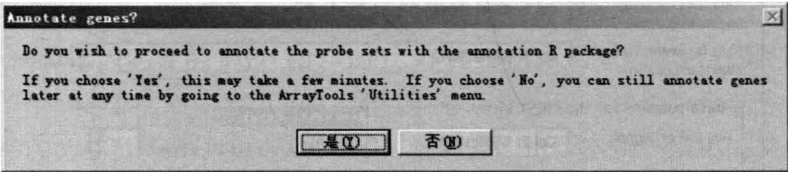


图2-13 选择是否对基因做注释

经过以上步骤，基因芯片的预处理已经结束，我们获得了经过过滤的标准化数据，基于该数据可以进行接下来的深入分析。

第三节
特征基因挖掘

Section 3 Feature Gene Mining

特征选择(feature selection)是按照某一个评价准则从描述实例的高维特征集中搜索出低维的最优特征子集,从而最大限度地提高分类器的性能(分类器将在本章第五节详细介绍)。特征个数越多越容易引起“维度灾难”,得到的模型越复杂,可移植性也随之下降。特征选择能够通过剔除冗余特征、减少特征数目,达到提高模型精度、减少运行时间的目的。基因芯片数据具有高维度、高信噪比、高相关(冗余)的特点,基于基因表达谱从大量的基因中寻找对疾病有鉴别力的基因或疾病相关基因作为疾病标志物,也就是基因芯片分析中的特征选择问题。

一、特征选择的过程

特征选择过程通常包含以下四个方面: 开始点的选择; 搜索策略(search strategy); 评价准则(evaluation criteria); 停止条件。在实际学习中,选择一个较优的特征子集主要依赖于搜索策略和评价准则这两个方面。总的来说,特征选择的理想效果是: 将所有可能的特征子集作为属性训练分类器,然后选取能够使分类器达到最佳分类效能的特征子集。

(一) 开始点的选择

从大量特征中选择构成最优特征子集的特征,需要对特征全集进行搜索,搜索的方向受开始点选择的影响。如果把空集作为初始特征子集,逐次递加特征进入特征子集,称为向前选择(forward selection); 如果把含有所有特征的特征全集作为初始特征子集,逐次剔除特征称为向后选择(backward elimination)。此外,还有比较复杂的开始点选择方式,例如: 如果把包含一定数目特征的特征子集作为初始特征子集,然后向外扩展,称为双向搜索(bi-direction search); 从随机选择的特征子集开始搜索,再随机增加或减少特征,则称为随机搜索(non-deterministic search)。

(二) 搜索策略

理想情况下,搜索策略应该能够以较低的计算花费找到最优特征子集。但通常这两个条件不能同时满足,需要折中权衡。现有的搜索策略有许多种,按照寻找特征子集的过程大

致可归纳为以下三类：完全搜索（complete search）、启发式搜索（heuristic search）和随机搜索（non-deterministic search）。

1. 完全搜索 穷举法是最常见的一种完全搜索方法，其通过遍历所有可能的特征子集，从而保证搜索到的是最优特征子集。分枝定界法是按照定好的界进行分枝，其本质仍是穷举法。将所有特征子集组织成树形结构，这些特征子集构成解空间。所谓"分枝"就是不断将解空间分割成更小的解子集，"定界"则是对分割得到的更小解子集计算一个上界或下界，对超越出该界限的解子集（即非最优的特征子集）不再进行分枝。通过缩小搜索范围，分枝定界法能够提高算法效率，同时也能够求得最优解。以上两种方法都是以大量的时间和空间消耗为代价来获得最优解，当特征维数较小时，可以获得很好的效果。但当特征维数较大时，由于运算量太大，用计算机实现也可能会遇到困难。因此，当特征维数较大时，可以采用启发式搜索来缩小搜索范围，获得次优解。

2. 启发式搜索 贪婪登山法（greedy climbing hill）、遗传算法（genetic algorithm）、模拟退火算法（simulated annealing algorithm）和Tabu搜索算法等都属于启发式搜索算法。这些算法的主要思想是人类经过长期对物理、生物和社会的仔细观察和实践，通过对这些现象的深刻理解，逐步向自然学习，模仿它们的运行机制而得到的，如模拟退火受物理学上固体物质的退火现象启迪，遗传算法则得益于生物进化论。启发式搜索通常从可行的初始解出发，采用迭代改进的策略，能较快接近最优解，但不能保证得到的解一定是最优解。

（1）贪婪登山法：贪婪登山法包括顺序前进法、顺序后退法和增r减q法。顺序前进法每次从未入选的特征中选择一个，加入已入选特征的集合，使其与已入选特征组合在一起时所得的目标函数最大（如分类器的正确率最高），直到特征数增加到一定数目为止。类似地，顺序后退法从所有特征构成的集合开始，每次剔除一个特征，所剔除的特征应该使仍然保留的特征所得到的目标函数最大。顺序前进法和顺序后退法都是进行简单的串行搜索，会遗漏掉大量的特征组合，而且一旦某个特征被选入（剔除），就不能再剔除（选入）。为了弥补这种不可回溯的缺点，可以在搜索过程中加入局部回溯，这就是增r减q法。增r减q法是在k步中先用顺序前进法逐个加入特征到$k+r$个，然后再用顺序后退法逐个剔除q个特征。

（2）遗传算法：根据达尔文的生物进化论，自然界中的每个个体不断对环境学习和适应，然后通过交叉方式产生新的后代，这就是基因的遗传。通过遗传，这些后代继承了双亲的优良特性，并继续对环境学习和适应。基因突变发生在交叉之后，有利的变异由于自然选择的作用得以遗传与保留，而有害的变异则将逐步被淘汰。遗传算法是一种模拟生物的进化过程（遗传、变异和自然选择）的用于优化的搜索算法，可以避免出现局部极值。遗传算法遵循适者生存、优胜劣汰的法则，即在寻优过程中将有用的特征保留，去除冗余或不相关的特征。由于遗传算法具有良好的并行性、通用性和稳健性，因此它在特征选择领域具有广阔的应用前景。

下面介绍遗传算法用到的基本术语：

基因链码：自然界的生物所表现出的各种性状是遗传基因所决定的，生物的遗传特性使生物界的物种能够保持相对的稳定。使用遗传算法解特征选择问题时，需要把问题的每一个解编码成一个基因链码。基因链码是对多个基因编码所得到的字符串，字符串的每一位代表一个基因。一个基因链码就代表问题的一个解，相当于自然界中的个体。简单的编码可以采用二进制形式，即用N位的0，1代码构成的字符串表示一个特征集合，其中数字1对

应的特征被选中，数字0对应的特征未被选中。

群体：许多个体的集合构成群体。个体代表问题的一个解，群体则是问题的多个解构成的集合。

交叉：选择群体中的两个个体作为双亲，在配对的两个个体中设置截断点，然后交换两个个体的信息产生后代个体。简单的交叉方法可以通过随机配对产生双亲个体（对应两个基因链码）并随机选择截断点（即基因链码中的某一位），然后将两个基因链码在截断点切开并交换其后的基因链码，从而组合成两个新的基因链码，也就是双亲个体通过交叉繁殖出同样数量的后代个体。复杂的交叉方法可以自行设定截断点和双亲个体，无论使用何种交叉方法都以培育出更适应环境的后代为最终目的。

变异：这里的变异沿用了生物学中基因突变的概念，生物的变异特性使生物个体产生新的性状，最终积累形成新的物种。变异方法是针对某个个体（即一个基因链码）随机选取某个基因（即基因链码的某个位点），将该基因进行变异操作。比如，对二进制编码得到的基因链码，只需将已选的位点处的数字从1换成0或者从0换成1。

适应度：每个个体对应优化问题的一个解，根据优化问题的目标函数，对应每个解求得函数值。如果优化问题要求取最大，那么使函数值越大的解越接近最优解，也就是表明该个体对环境的适应度越高。

（3）Tabu搜索算法：Tabu搜索算法是一种全局逐步寻求最优的算法，该算法假定一个解的邻域中往往存在性能更好的解。Tabu搜索算法应用于特征选择中时，解的性能高就是指使用相应的特征子集（或特征组合）的分类效果好，一般用可分性判据来度量。该算法中使用了“集中”和“扩散”两个策略，局部搜索过程体现“集中”的思想，也就是从一点出发，在这点的邻域内寻求性能更高的解，达到局部最优解而结束。“扩散”的思想则体现在跳出局部最优的过程，通过设置Tabu表来实现跳出局部极小。Tabu表用来记录近期搜索过的解，如果一个解在Tabu表中，则说明近期该解曾被访问过，在未来一段时间内禁止访问该解，这种解被认为处于休眠状态。Tabu表越长，搜索的范围越广泛，获得性能较高的解的可能性越大。Tabu算法通过禁止访问Tabu表中已记录的解而实现对邻域之外更大区域的搜索，最终能够跳出局部最优找到性能更高的解。在一些情况下，需要将Tabu表中处于休眠状态的解激活，使其再次参与搜索过程。Tuba算法的过程决定了得到的最终解是在所有搜索过的解中的最优解。

（4）模拟退火算法：模拟退火算法得益于对统计物理中固体物质的结晶过程的研究。固体物质内部粒子的不同结构对应于粒子的不同能量水平。在高温条件下，粒子的能量较高，可以自由运动重新排序，在低温条件下，粒子能量较低。在升温过程中，固体物质内部的粒子随温度升高变为无序状态，能量增大。从高温开始缓慢降温的过程称为“退火”。在退火过程中，粒子逐渐趋于有序，粒子在每个温度都能达到热平衡状态。当固体物质被完全冷却时，最终形成处于低能状态的晶体。

模拟退火是模拟物理系统退火过程的随机迭代寻求最优的算法，理论上具有一定概率下的全局优化性能。假设要解决一个寻找最小值的优化问题，在迭代的开始阶段，搜索过程的随机性很大，除了接受优化解（使目标函数值变小的解）之外，还以由温度相关的系数来控制的概率接受恶化解（使目标函数值增大的解）。当迭代一定次数后，进入下一个迭代阶段，此时算法接受恶化解的概率较前阶段要有一定降低。如此不断迭代，直到达到停止准则时

算法终止，在最后阶段算法将不接受恶化解。模拟退火算法可以避免陷入局部极小值，得到的是一个优化解，并且它可能是全局最优解，但是不能保证它一定是全局最优解。模拟退火算法在应用于特征选择中时，首先要给出初始温度和初始特征子集，然后给出该特征子集的邻域和温度下降方法。

3. 随机搜索　与完全搜索和启发式搜索不同，随机搜索以随机的方式搜索下一个特征子集，当前的子集不是根据某个决策规则直接增长或缩小得来的。该搜索策略计算复杂度较高，但是通过设置迭代次数在一定程度上可以降低复杂度。

(三) 评价准则

评价准则通常分为独立标准和非独立标准。

1. 独立标准　距离、相似性(或相关性)和互信息等多种测度都可以作为独立标准。明氏距离是常用的距离测度，明氏距离通过考查基因表达值向量的距离大小来反映基因表达的差异。理想的分类效果是类内基因之间的距离很小，而类间基因之间的距离很大。相关系数常常作为评价相似性的测度，基因与类别的相关系数反映基因与类别之间的相关程度。选择的特征基因与类别的相关程度应大于其他基因与类别的相关程度。相关系数可以分为线性相关，如皮尔森相关系数(pearson correlation coefficient)和非线性相关，如斯皮尔曼秩相关系数(spearman's rank correlation coefficient)。此外，信息论中的互信息指标也可用来评价基因与类别的相关程度，互信息越大说明该基因的表达模式与类别越相关，以互信息最大化为标准可以用来评价特征基因的优劣(计算公式请参考本章第四节)。

2. 非独立标准　以分类正确率为准则的标准属于非独立标准。在有监督学习中，分类的主要目标是分类器预测正确率最大化。因此，可利用分类器的预测正确率作为特征选择的评价标准。另外也可考虑其他的指标，如泛化能力和时间复杂度。这里的泛化能力是指利用一种分类器选择出的特征基因适合用于其他分类器的能力。

(四) 停止条件

由于所有特征构成的可能的特征子集的数量很大，考察所有的特征子集通常不可实现，因此需要某种停止搜索的条件，例如：迭代次数、特征子集评价标准达到阈值或不再继续提高等。一个较优的停止条件是到达搜索终点时，选择的子集为最优特征子集。

二、特征选择方法的分类 》》》

目前，特征选择方法主要分为三类：过滤法(filter method)、缠绕法(wrapper method)、镶嵌法(embedded method)。在过滤法中，特征选择过程独立于分类算法，利用一些独立的评价标准预先完成特征子集选择，然后再进行分类器的归纳学习，该方法通常是对单基因进行逐一评价，如统计检验、互信息等。过滤法在应用过程中鲁棒性强，运行速度快。缠绕法中特征选择过程与分类算法绑定，将分类算法的效能作为入选特征基因子集的评价准则，由于选择的特征基因子集能够与分类器的决策机制很好地吻合，因此，对检验样本的划分可获得较高的准确率。镶嵌法则是特征选择过程与分类过程并行的一类特殊方法，在构建分类器的同时进行特征基因选择，决策树算法是常见的镶嵌式特征选择方法。

（一）过滤法

过滤法是机器学习中进行特征选择最早使用的方法，所有的过滤法都只依赖数据本身的内在结构信息而不依赖分类算法对特征子集的评价，不考虑所选的特征子集对分类器性能的影响，也就是说，在分类算法运行之前进行特征选择，二者相互独立。我们可以根据需要选择特征集合，比如，使特征之间的相关度尽可能低，此种方法适合较大的数据集。过滤法由于与分类算法相互独立，不考虑特征子集对分类器分类效能的影响，所选出的特征子集分类性能一般弱于缠绕法。

*t*检验法　该方法运用统计学上传统的*t*检验寻找在两类间特征值有差异的特征，应用于基因表达谱分析中，就是寻找疾病和正常状态之间差异表达的基因，这些基因就是特征基因。*t*检验法首先计算每个基因*i*的统计量t_i:

$$t_i = \frac{\overline{x_{i1}} - \overline{x_{i2}}}{\sqrt{\frac{s_{i1}^2}{n_1} + \frac{s_{i2}^2}{n_2}}} \tag{2-5}$$

其中$\overline{x_{i1}}$和$\overline{x_{i2}}$分别表示基因*i*在第一类和第二类样本中表达水平的均值，s_{i1}和s_{i2}分别是第一类和第二类样本中基因*i*表达水平的标准差，n_1和n_2分别是两类中样本的数目。然后根据统计量t_i得出相应的假设检验的概率p_i值。由于对涉及的多个基因进行了多次假设检验，Ⅰ类错误率上升，所以要对所得的*p*值进行多重检验校正。常用的多重检验校正方法有Bonferroni校正、FDR等。

（二）缠绕法

缠绕法依赖于特定分类器，将分类算法嵌入特征选择过程中，以分类正确率为评价准则，通过一定的搜索策略识别优化的特征基因子集。缠绕法计算量较大，适合较小的数据集。与过滤法相比，由于特征选择的结果由分类器的正确率来评价，因此，缠绕法能够将所选的特征与分类器的决策进行较好地结合，通常可以实现分类准确率最大化。

遗传算法与支持向量机耦合的特征选择方法（genetics algorithm-support vector machine，GA-SVM）是一种典型的缠绕法，其采用遗传算法作为搜索策略，以支持向量机分类器的效能作为特征子集优劣的评价准则。该算法是个递归的过程，对亲代进行遗传操作产生后代。在这种方式下，优良的特征基因子集（即提高SVM分类正确率的特征基因子集）不断被“进化”，直到遇到停止条件。其中，特征基因子集的编码、群体的初始化、适应度计算、遗传操作、控制参数的设定（群体大小，最大迭代数等）是GA-SVM的核心内容。此外，选用不同的搜索算法和分类算法，缠绕法特征选择方法还有许多，如基于遗传算法和*k*近邻耦合的特征选择方法（genetic algorithm and the *k*-nearest neighbor，GA-KNN）、支持向量机-递归特征消除法（support vector machines-recursive feature elimination，SVM-RFE）等。

Li等人利用GA-SVM方法分析了弥漫性大B细胞性淋巴瘤（diffuse large B-cell lymphoma，DLBCL）相关的基因芯片数据，识别区分DLBCL两个亚型生发中心B细胞DLBCL（GCB-like DLBCL）和活化B细胞DLBCL（AB-like DLBCL）的特征基因子集。基因表达谱包括21个GCB-like DLBCL样本和21个AB-like DLBCL样本，以及4026个经过预处理的基因。在该研究

中,特征子集被编码为二进制串,串中每个位置的1/0表示该位置代表的基因在特征子集中出现/不出现。每个特征子集表示一个个体,初始群体的大小设为40,在随机生成初始群体时,为了保留较多具有分类信息的基因,初始群体中每个个体包含大约1/2的全部基因。以每个个体所包含的基因为特征构建支持向量机分类器,以分类正确率作为适应度指标评价个体的好坏。该方法通过生存竞争实现特征子集的优化,首先为避免每一代中的最优解丢失,保留前50%的优良个体直接进入下一代;然后,通过随机进行交叉和变异产生下一代的另外50%的个体。为了找到具有代表意义的较小的特征基因子集,因此采用逐步缩小特征基因数目的方法:在上一轮执行完成的基础上把最优个体对应的表达谱子矩阵作为新的研究对象,重复执行上述过程进行迭代,直到分类的准确率下降小于0.001且选出的特征数不再变化为止。该研究共迭代了12次,特征基因数目的变化为:4026,1995,984,504,256,132,70,41,25,18,13,7,最终得到了由7个基因(*CYSLTR1*,*MME*,*D13S2489E*,*PIK3CG*,*SHMT2*,*Hs.348293*,*Hs.291994*)组成的优化的特征基因子集。该方法的流程图如图2-14所示。最后,作者又将GA-SVM方法与其他的特征选择方法(*t*检验、GA-KNN等)进行了比较,在多种分类器下,GA-SVM选取的特征子集的分类贡献均高于其他的基因子集(图2-15)。

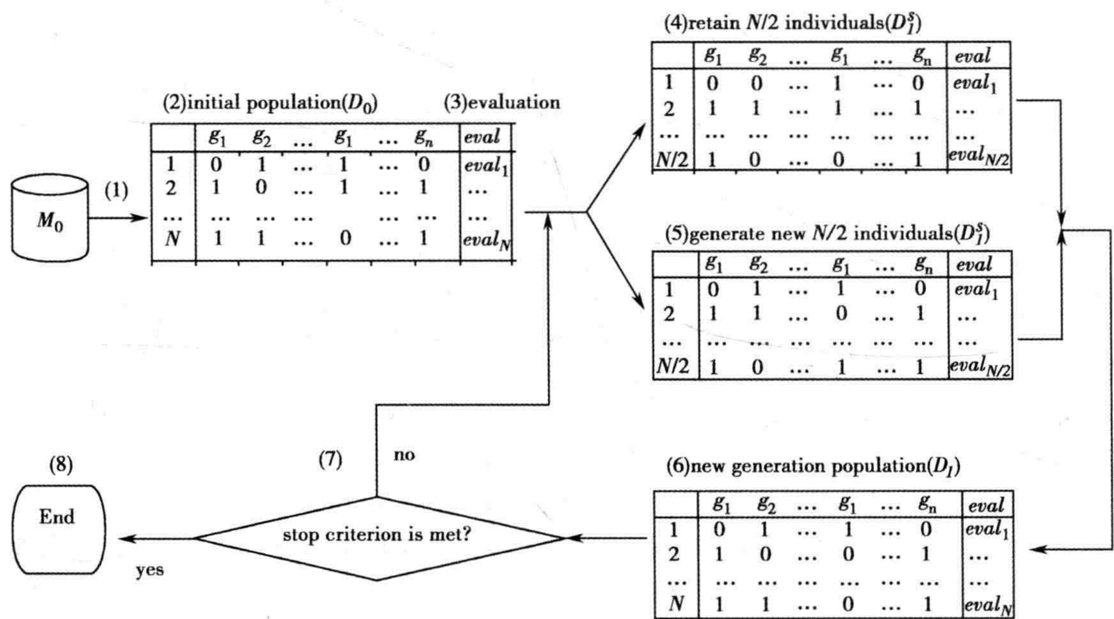


图2-14 GA-SVM算法流程图

(三)镶嵌法

镶嵌法利用既有分类功能又有特征选择功能的分类算法,将分类和特征选择同时进行。以决策树算法为例,这种算法用树型结构来表示分类规则。决策树的构建就是进行特征选择的过程。使用决策树进行决策的过程就是从根节点开始,测试待分类样本中相应的特征属性,并按照属性值选择输出分支,直到到达叶子节点为止,将叶子节点所代表的类别作为最终样本的分类,在分类器构建的过程中,每一个非叶子节点上用于对样本进行分类的基因组合起来就构成了特征基因子集。常用的决策树算法有CART、ID3和C4.5等。

Li等人提出了一种基于递归决策树特征基因选择的集成方法,在构建决策树分类器的

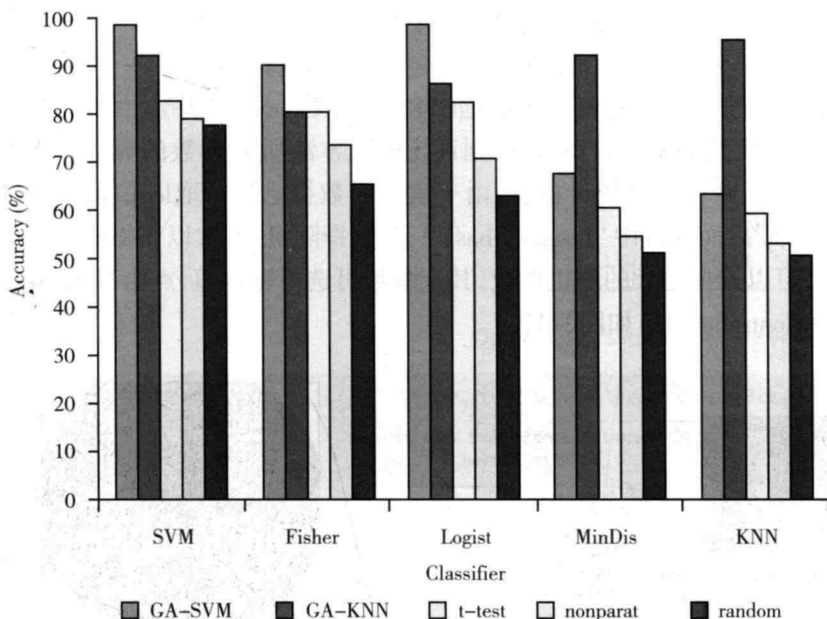


图2-15　特征选择方法的比较

同时进行特征选择，该方法成功应用于结肠癌基因芯片（40个肿瘤组织和22个正常组织中的2000个人类基因的表达水平）数据分析中。该方法的基本思路是首先分别将结肠癌和正常组织样本随机分为近似大小的5个不重叠子集，即结肠癌子集D_i（i=1，2，…，5）和正常组织子集N_i（i=1，2，…，5），D_i和N_i的一个随机配对构成一个检验集，剩余的所有样本构成一个训练集，这样一次抽样可产生25个训练集和检验集对，重复该过程20次，共获得500对训练集和检验集对。在每对数据集上执行一次特征基因识别过程：基于训练集构建决策树分类器，每个非叶子节点上的基因构成特征基因子集$G_d=\{g_1^d, g_2^d, \cdots, g_k^d\}$（$d$= 1，2，…，500），并利用检验集进行分类效能评价。这样就得到了一系列特征基因子集G_1，…，G_d，…，G_{500}，然后计算每个基因在所有这些子集中出现的加权频率值（权值可定义为分类器的正确率）FV，为了得到具有统计学显著性的特征基因，作者将样本的类别进行随机扰动，重复上述过程，计算在随机情况下基因的加权频率值FV^0，构建随机分布，对应于显著水平0.01的经验阈值0.035，记作FV^0_β = 0.035（β= 0.01），保留那些FV值大于FV^0_β的基因作为特征基因，最终共识别出20个基因构成了优化的特征基因子集。该研究的结果表明基于决策树的特征基因选择方法能够有效识别复杂疾病相关基因。

三、应用举例

基因芯片数据分析中常见的差异表达基因识别方法可以归为特征选择中的过滤法。基因芯片显著性分析（significance analysis of microarray，SAM）是目前使用较为广泛的差异表达基因分析软件。SAM通过计算每个基因的统计量D值，寻找对疾病有鉴别力的基因。SAM是一个Excel的插件，安装成功后以加载项的形式出现在Excel菜单栏中，在此，以上面已经进行过预处理、标准化的基因芯片数据GSE5281中的内侧颞回区域的基因表达谱为例介绍该软件实现差异表达分析的过程。由于这是一个两类非配对样本的问题，因此，应当以如下格

式准备数据文件：第一行为类别标签，1代表正常，2代表疾病；第一列是基因Entrez ID，第二列是基因Symbol，其余列为表达值；还需要注意前两列首行应为空。

首先，要将经过预处理的表达谱数据GSE5281用Excel打开并选中所有数据，在Excel菜单栏的加载项中找到SAM，运行SAM得到设定SAM方法所需参数的界面，如图2-16。这里我们选择两类非配对样本做统计检验，由于表达谱数据已进行取log值的处理，因此在"Are data in log scale？"后面要选中"Logged（base 2）"。选择随机100次以获得统计量D相应的p值，按照不同需要可以选择更大的随机次数，其余参数可选择默认值，点击"OK"即可继续运行，弹出SAM Plot Controller窗口如图2-17。

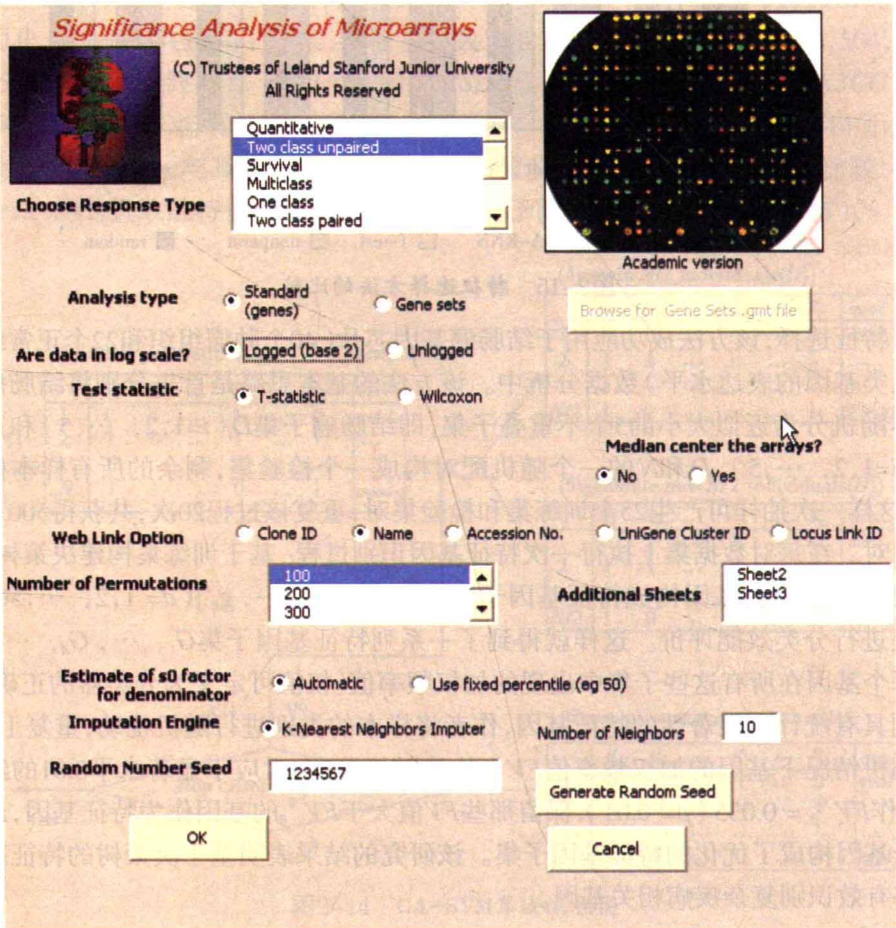


图2-16　SAM的参数设定

接下来，在SAM Plot Controller窗口通过设定Fold Change值和delta值来控制差异表达分析的结果。点击"List Delta Table"可以获得delta值与FDR值的对应关系。在此，我们找到FDR为0.01时对应的delta值为0.68。然后，通过滑动滑块或手动输入delta值，点击"List Significant Genes"就得到了FDR小于0.01的差异表达基因，本例共选出了2209个在阿尔茨海默病病人和正常人脑组织中表达发生显著改变的基因。此外，SAM还以图形化方式"SAM Plot"对结果进行展示（图2-18），其中显示了差异表达基因的期望得分与观察得分的关联关系，上调基因用红色标识，下调基因用绿色标识。

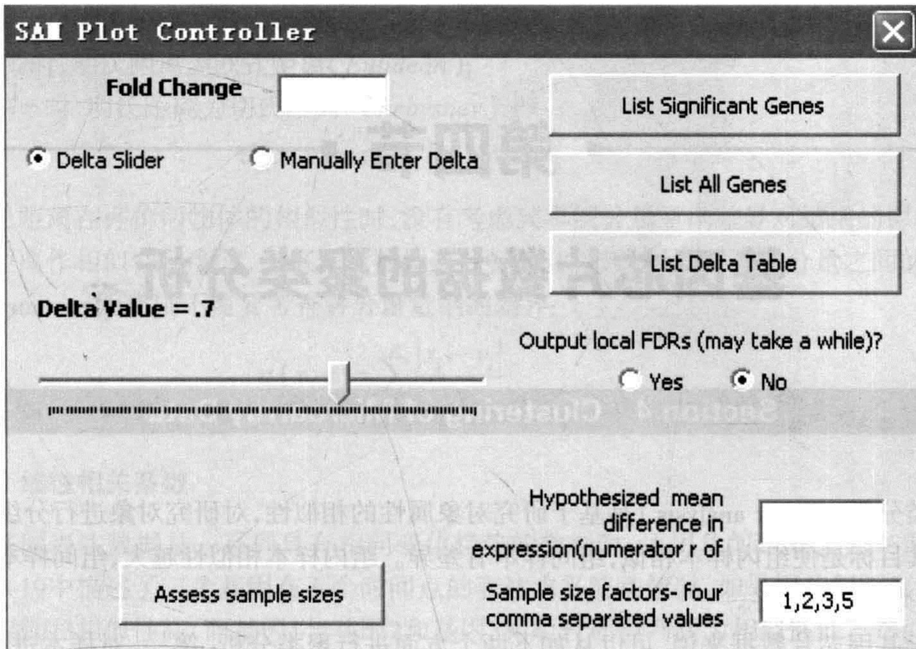


图2-17 SAM Plot Controller窗口

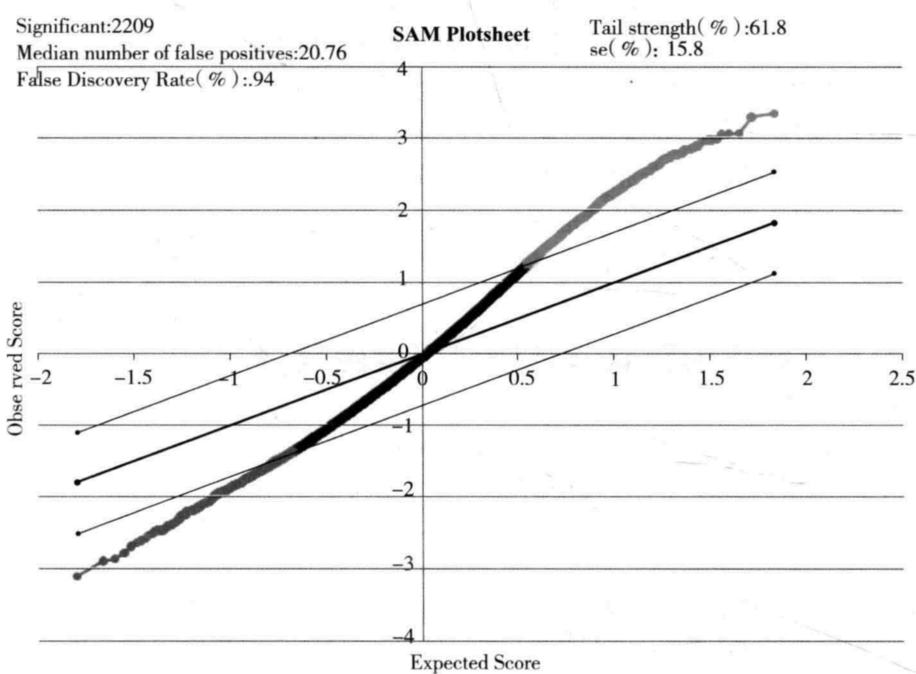


图2-18 SAM Plot

第四节

基因芯片数据的聚类分析

Section 4 Clustering of Microarray Data

聚类分析(cluster analysis)是基于研究对象属性的相似性,对研究对象进行分组的一种方法。其目标是使组内样本相似,组间样本有差异。组内样本相似性越大,组间样本差异越显著,聚类效果就越好。

对于基因芯片数据来说,可以从如下两个方面进行聚类分析:第一,对样本进行聚类,即研究对象为样本,属性为基因,基因表达相似的样本聚为一类。在人类癌症的研究中,对样本进行聚类主要应用于癌症亚型的识别,由于肿瘤异质性的存在,即临床表型相同的肿瘤往往具有不同的分子机制,因此利用基因芯片数据对肿瘤样本进行聚类,有助于从分子层面识别肿瘤新的亚型,为肿瘤患者的个性化诊疗提供重要参考。第二,对基因进行聚类,即研究对象为基因,属性为样本,在样本空间中表达模式相似的基因聚为一类,同一类的基因往往具有功能上的一致性,即参与相同的代谢通路或者编码同一个蛋白质复合物等。

聚类分析中最主要的两个要素是评价研究对象属性相似性程度的距离(或相似性)尺度(distance scale)和将研究对象分组的聚类算法(clustering algorithm)。

一、聚类分析中的距离(相似性)尺度函数

距离(相似性)尺度函数是评价研究对象相似性程度的函数。常用的表达相似性尺度有几何距离、线性相关系数、非线性相关系数和互信息等。

(一) 几何距离

几何距离可以衡量研究对象在空间上的距离,空间上相近的物体可以运用几何距离判断为同一类,而空间上较远的物体判断为不同类。

常见的几何距离是明氏距离:

$$d(x,y)=\left\{\sum |x_i-y_i|^{\lambda}\right\}^{\frac{1}{\lambda}} \tag{2-6}$$

其中x和y为样本向量或基因向量,xi和yi为对应的第i个分量,明氏距离通过考查各分量的差异来衡量两物体的距离大小。

当$\lambda=1$时，明氏距离为马氏距离（*Manhattan*）；

当$\lambda=2$时，明氏距离为欧式距离（*Eulidean*）；

当$\lambda=\infty$时，明氏距离为切氏距离（*Chebyshev*），即

$$d(x,y)=\max_i |x_i-y_i| \tag{2-7}$$

明氏距离在评价两物体的相似性时，没有考虑到不同分量量纲差异对结果的影响，所以用明氏距离作相似性尺度时，应该先对数据进行标准化处理，以消除不同分量之间的量纲差异。而Camberra距离则不需要考查各分量量纲的差异：

$$d(x,y)=\sum_{i=1}^{p}\frac{|x_i-y_i|}{|x_i+y_i|} \tag{2-8}$$

（二）线性相关系数

当基因表达数据是一系列具有相同变化趋势的数据时，运用几何距离会丢失重要的信息。图2-19中描述了三个基因在五个时间点的表达水平波动情况，如果用几何距离衡量，则基因2和基因3相似性高，而基因1与基因2和基因3相距较远会判断为相似性低。然而，基因1的表达水平在不同时间点与其他两基因具有相似的波动趋势和波动幅度，通常这种在不同时间点或样本中表达模式相似的基因也有可能具有功能上的相关性，但是用欧氏距离可能

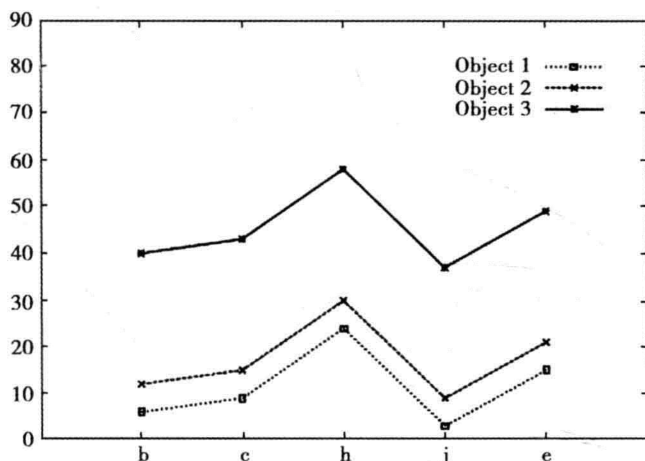


图2-19 三个基因在五个时间点的表达值波动图

就会忽略这种具有生物学意义的基因相关关系。

这时，一般采用皮尔森相关系数（pearson correlation coefficient）来衡量基因表达模式的相似性。公式如下：

$$r=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma_x}\right)\left(\frac{y_i-\bar{y}}{\sigma_y}\right) \tag{2-9}$$

其中$\bar{x}$为基因向量x的期望值，σ_x为x的标准差；$\bar{y}$为基因向量y的期望值，σ_y为y的标准差，n为向量x的维数，即时间点或样本的个数。

(三) 非线性相关系数

在实际问题中可能存在这样的基因,它们在表达上不具有严格的线性相关关系,但是波动趋势却相同,即具有同升或同降的变化趋势。在这种情况下可以用非线性相关模式来衡量基因间的距离。

非线性相关模式一般用斯皮尔曼秩相关系数(spearman's rank correlation coefficient)进行衡量:

$$\gamma = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)} \tag{2-10}$$

其中d为每对观察值xi与yi的秩次之差,n为时间点或样本的个数。

(四) 互信息

线性与非线性相关系数都只能衡量基因间的单调相关关系,而对于那些在前阶段正相关(负相关)、后阶段负相关(正相关),即具有非单调性特点的两个基因来说则不适用。对于这种相似关系,可以用互信息来度量:

$$\gamma = H\left(x\right) - H\left(x \mid y\right) \tag{2-11}$$

其中$H(x)$表示x的熵,$H(x|y)$表示x的条件熵。当x和y为离散型向量时,条件熵的计算方式表示如下:

$$H\left(x \mid y_J\right) = -\sum_{I=1}^{n} p\left(x_I \mid y_J\right) \log p\left(x_I \mid y_J\right) \tag{2-12}$$

$$H\left(x \mid y\right) = -\sum_{I=1}^{n}\sum_{J=1}^{m} p\left(y_J\right) p\left(x_I \mid y_J\right) \log p\left(x_I \mid y_J\right) \tag{2-13}$$

其中,p表示概率密度函数,可以由频数估计;n和m分别为离散化x和y时的离散化单位。在计算互信息时采用的离散化方式会造成一定的信息损失,一般离散化单位的估计由向量x和y的长度决定。

$$n \leqslant \log_2 size(x) \tag{2-14}$$

$$m \leqslant \log_2 size(y) \tag{2-15}$$

二、常用的聚类方法

(一) k均值聚类

k均值聚类(k-means clustering)是根据对象的均值进行聚类划分的分割算法,适用于各种数据类型,受初始化问题的影响较小,算法简单,运算速度快。

k均值聚类的具体分析流程如下:

1. 初始化类中心,随机选择k个初始质心,其中k是自定义参数,表示所期望簇(类)的个数。

2. 计算每个对象与质心的距离,将每个样本指派到距离最近的质心,指派到一个质心的对象组成一个簇。

3. 重新计算每个簇的样本均值,作为更新后簇的质心。

4. 重复2~3步,直到每个簇不再发生变化为止。

需要指出的是,在实际应用中大部分收敛都发生在早期阶段,通常使用较弱的终止条件结束该算法,因此,步骤4可改为"直到仅有1%的点改变簇"为止。

图2-20举例说明了k均值的聚类过程,从3个质心出发,通过4次指派和更新,找出最后的簇。其中,质心用符号"+"表示,属于同一个簇的所有点具有相同形状的标记。

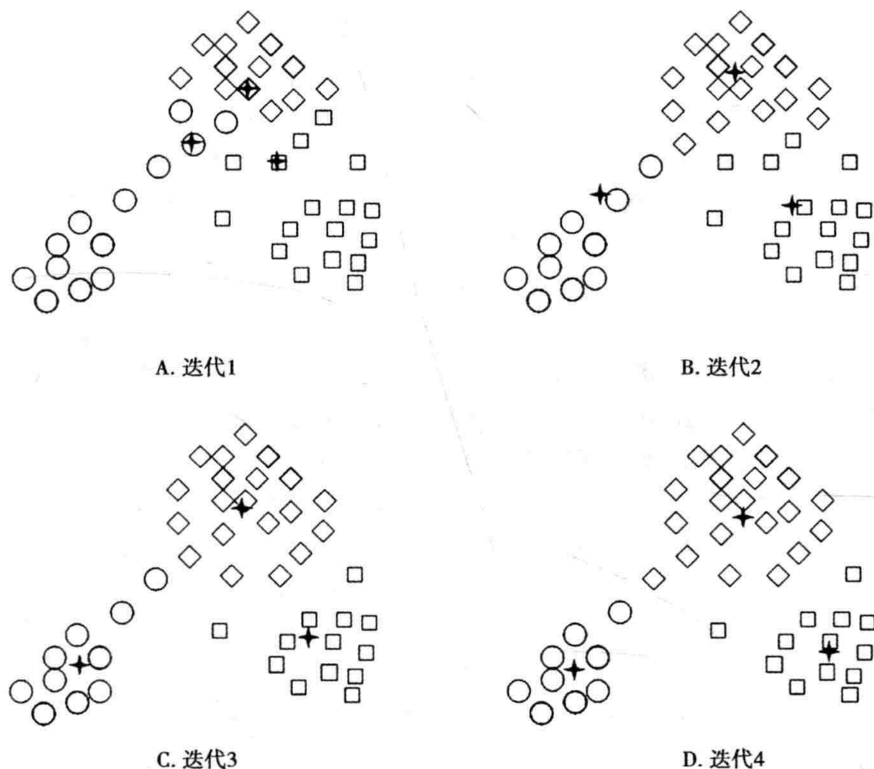


图2-20　k均值聚类过程示意图

k均值聚类可以看作是优化问题,其优化目标是使类内样本两两间的距离之和最小,通常表示为:

$$w(C)=\frac{1}{2}\sum_{c=1}^{k}\sum_{C(i)=C(j)=c} d_E\left(x_i,x_j\right)^2 \tag{2-16}$$

其中,k表示簇的个数,C表示簇的结构,x_i和x_j是属于同一个簇中的样本,$C(i)$和$C(j)$分别是样本x_i和x_j的类别,d_E表示两个样本的欧氏距离。

k均值聚类算法的结果依赖于初始质心的选取,不同的质心将可能产生不同的聚类结构。为了克服这个问题,可以采用多个初始化方式,选出具有最小$w(C)$的聚类结果作为最佳的类结构。另外,k均值聚类需要预先指定类别个数,但是由于是无监督学习方法,在实际应用中一般不知道真实的类别个数,一些启发式的方法可以帮助确定k的取值。例如,假设

有6个研究对象，则遍历6个对象可能的聚类类别数，计算各个情况下的$w(C)$，选择$w(C)$下降最快时的k值作为最佳类别数。

（二）层次聚类

层次聚类（hierarchical clustering）是另一种常用的聚类方法，常常使用系统树图（dendrogram）的方式表示，如图2-21所示：

层次聚类的方法可以分为两类，分别为凝聚法和分裂法。

凝聚法（agglomerative）是一种自下而上的聚类方法，从单个点作为个体簇开始，每一步合并两个最邻近的簇。

分裂法（divisive）是一种自上而下的聚类方法，从一个包含所有点的簇开始，每一步分裂一个簇，直到仅剩下单点簇为止。

目前，凝聚法层次聚类技术使用最为普遍，其计算步骤如下：

1. 计算邻近度矩阵。
2. 合并距离最近的两个簇。
3. 更新邻近度矩阵，以反映新的簇和原来的簇之间的相似性。
4. 重复过程2~3，直到仅剩下一个簇为止。

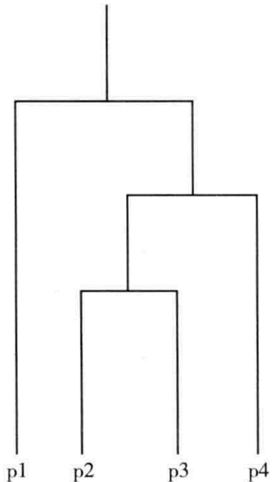


图2-21　层次聚类的系统树图

层次聚类算法的关键操作是计算两个簇之间的邻近度。常用的类间度量方法有：最小距离（single linkage）、最大距离（complete linkage）、平均距离（average linkage）和质心距离（centroid linkage）。如图2-22所示，最小距离以两类间距离最近的两点之间的距离作为两类的距离；最大距离以两类间距离最远的两点之间的距离作为两类的距离；平均距离则是遍历两类中所有两两点之间的距离，然后取平均值作为两类的距离；质心距离首先分别确定两类的质心，然后以质心间的距离作为两类的距离。

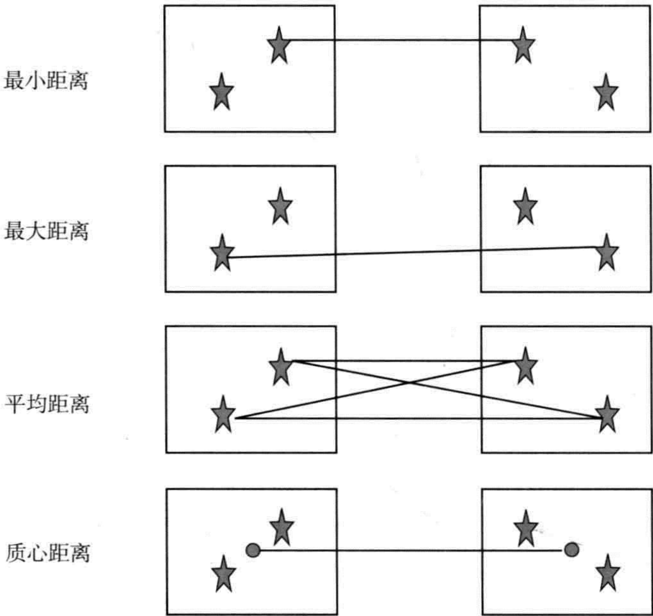
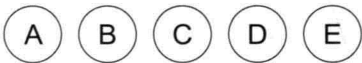


图2-22　类间相似性度量的方法

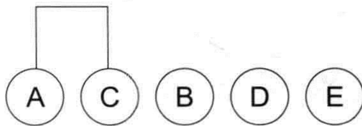
下面以一个例子说明自下向上的层次聚类算法过程，该实例采用欧氏距离衡量样本间的相似性。

1. 设有五个样本A、B、C、D、E，每个样本自成一类，运用欧氏距离计算它们两两之间的相似性，得出邻近度矩阵，此处为距离矩阵。



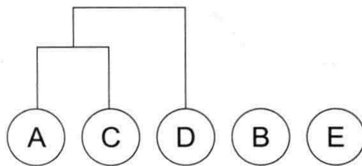
Dis	A	B	C	D	E
A		2	0.1	0.3	3
B			2.2	1.7	1.2
C				1.5	2.5
D					1.9

2. 由于A与C样本的距离最小，最先合并A与C样本。



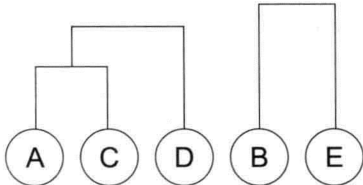
Dis	A	B	C	D	E
A		2	0.1	0.3	3
B			2.2	1.7	1.2
C				1.5	2.5
D					1.9

3. 合并后类别为三类，调整距离矩阵，即分别运用最小距离法计算B样本、D样本、E样本与AC类的距离，基于新的距离矩阵，合并AC与D样本。



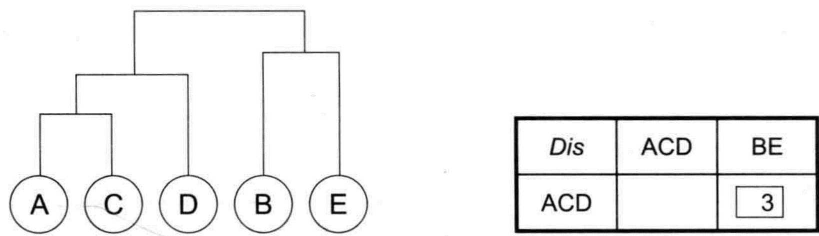
Dis	AC	B	D	E
AC		2	0.5	3
B			1.7	1.2
D				1.9

4. 继续调整矩阵，合并距离最近的B样本与E样本。



Dis	ACD	B	E
ACD		2.5	3
B			1.2

5. 合并ACD类与BE类，最后得到系统树图。



Dis	ACD	BE
ACD		3

三、应用实例

Cluster是聚类分析常用的软件之一，可以在多个平台上使用，包括Windows、Linux/Unix和Mac OS X。Cluster既提供图形用户界面，也可以输入命令行进行操作。下面以Cluster 3.0为例，介绍该软件的使用方法。

Cluster 3.0可以实现多种聚类方法，包括层次聚类、k均值聚类、自组织映射（self-organizing map，SOM）聚类，还可以实现主成分分析。这里以2209个在阿尔茨海默病和正常组织中差异表达的基因为属性，通过聚类将基因表达谱以热图（heat map）的方式显示出来。需要特别指出的是，聚类分析是无监督学习方法，通常是根据数据的内部特点将样本类别未知的数据进行归类，这里选择通过类别已知的样本识别的差异表达基因为例，仅仅是为了更好地展示数据，并介绍软件的使用步骤。

第一步，打开Cluster 3.0软件，界面如图2-23所示：

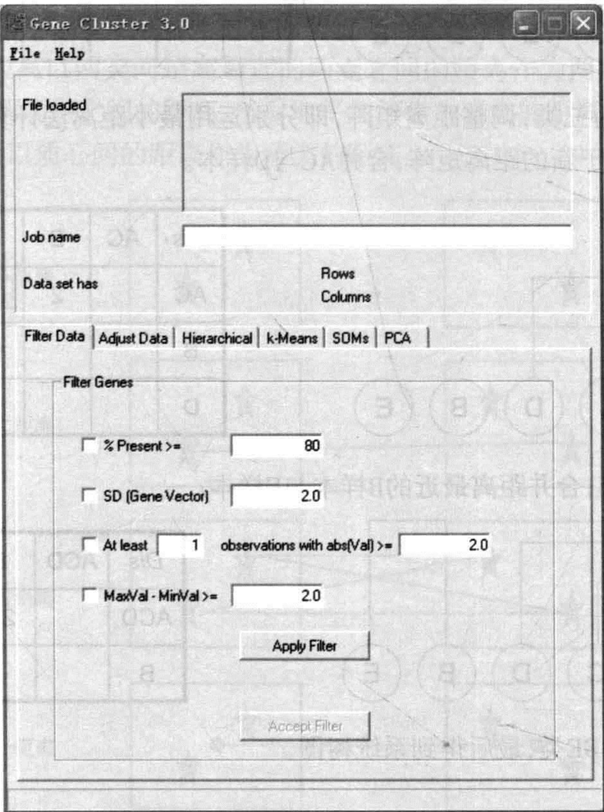


图2-23 Cluster 3.0界面

第二步，文件的导入

点击 “File” 按钮，选择 “Load data file” 选项，将前面章节中处理得到的GSE5281差异表达基因及相应的表达值导入Cluster 3.0中，结果如下（图2-24）：

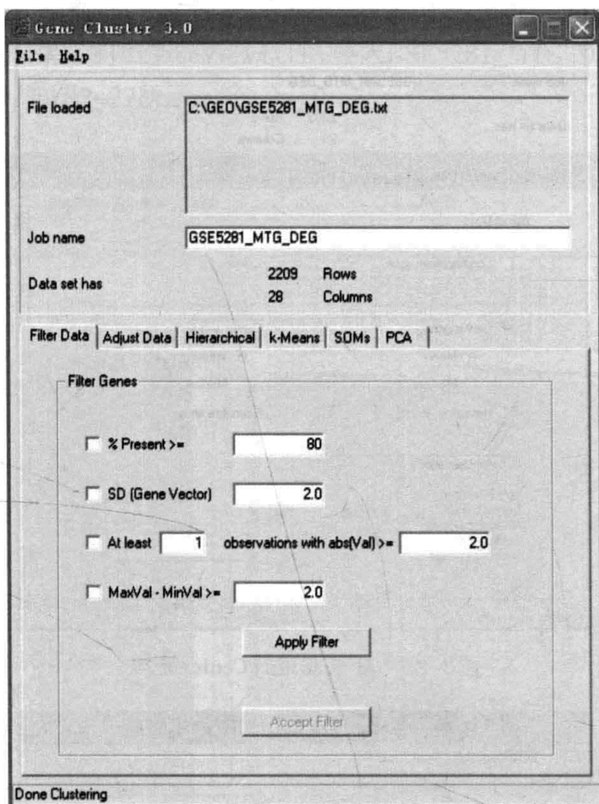


图2-24 Cluster 3.0文件导入

第三步，数据的校正

由于数据已经经过预处理，因此这里不再对数据进行过滤及标准化。为了更好地反映出各样本中基因表达值相对于基因平均表达值的高低，需要对数据进行Center处理，即将基因表达值减去其所在行、列的基因表达的均值或中值。点击 “Apply运行”。运行完毕后可以将运行结果保存。点击 “File” 按钮，选择 “Save data file” 选项，选择文件路径，可以将校正结果保存。具体操作如图2-25所示：

第四步，聚类分析

数据校正后，选择 “Hierarchical” 选项。该软件可以实现双向聚类，“Genes” 是对基因进行聚类，“Arrays” 是对样本进行聚类。这里在此对数据进行双向聚类，两种状态下都点选 “Cluster” 选项。下一步需要指定相似性矩阵（aimilarity Matrix）的计算方法。Cluster 3.0提供八种相似性矩阵的计算方法，包括皮尔森相关系数（pearson correlation coefficient）、斯皮尔曼秩相关系数（spearman's rank correlation coefficient）、欧式距离等，这里我们选择默认值 “Correlation (uncentered)”，即用皮尔森相关系数来计算相似性矩阵。最后，需要选择类间距离的度量方法（图2-26）。该软件给出四种度量方法，分别是质心距离、最小距离、最大距离和平均距离，在此我们选择质心距离的方法对数据进行聚类，参数选择如图2-26所示：

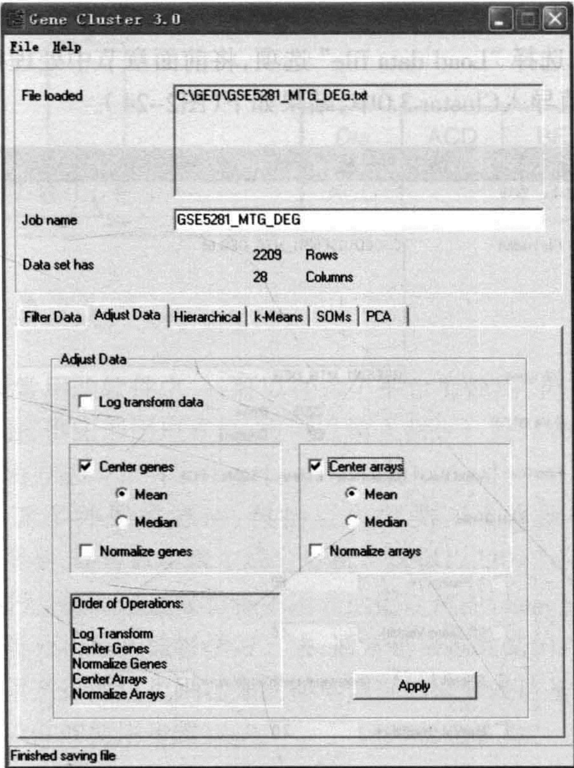


图2-25　对数据进行Center处理

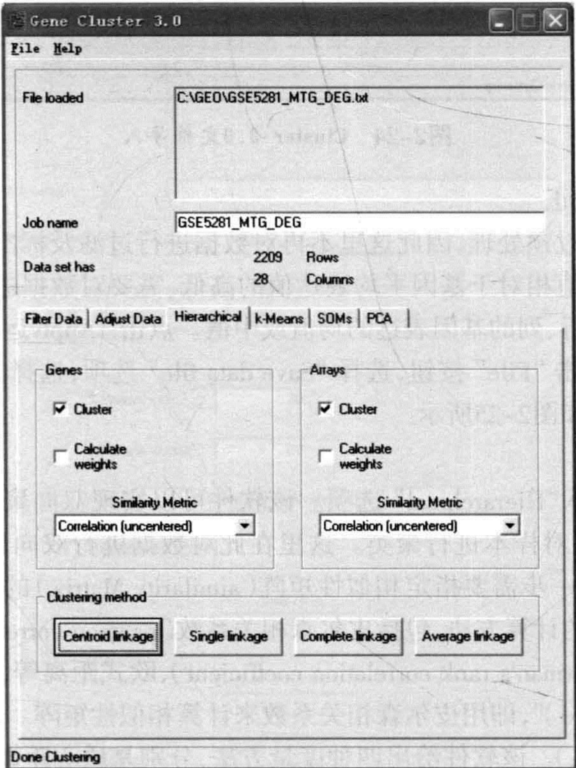


图2-26　对数据进行双向聚类

第五步，用TreeView查看聚类结果

聚类运算完成后会产生三个文件，GSE5281_MTG_DEG.gtr，GSE5281_MTG_DEG.atr，GSE5281_MTG_DEG.cdt。其中，GSE5281_MTG_DEG.gtr，GSE5281_MTG_DEG.atr分别记录对基因聚类和对样本聚类的过程，GSE5281_MTG_DEG.cdt记录聚类的结果。可以用TreeView可视化聚类结果。打开TreeView软件，导入GSE5281_MTG_DEG.cdt文件，即可得到聚类结果的树形结构，如图2-27所示：

图2-27 TreeView可视化聚类结果

第五节

基因芯片数据的分类分析

Section 5　Classification of Microarray Data

基因芯片数据的分类分析是一种有监督的学习方法，即样本的类别是已知的，通常以基因为特征，通过已知类别的样本训练分类器，评价分类器的效能，并对未知类别的样本进行预测。此外，为了提高分类器的分类效能，通常并不是用基因芯片上的所有基因来训练分类器，而是先进行特征选择，筛选出对分类有重要作用的特征基因子集（参考本章第三节），基于特征基因构建分类器。目前常用的分类方法包括线性判别法、贝叶斯分类法、人工神经网络、k近邻分类法、支持向量机、决策树和决策森林等，本节主要介绍k近邻、决策树、支持向量机等分类方法以及一些常用的分类效能的评价指标。

一、k近邻分类法

k近邻（k-nearest neighbor）分类的基本思想：对于给定的一个待分类的样本x，首先寻找与x最接近的或者最相似的k个已知类别的训练样本，然后根据这k个样本的类别标签来确定样本x的类别。

k近邻分类的具体步骤为：

1. 选取已知类别标签的训练样本集x。

2. 设置k（k为奇数）的初始值。k值的选取没有统一的方法（需要根据具体问题选择适当的k值）。常用方法是先确定一个初始值，然后通过不断地调试选择最优k值。

3. 在训练样本集中选出与待分类样本x最近的k个样本。常用的方法是计算已知类别样本和待分类样本间的欧式距离，选取与样本x距离最近的k个样本。

4. 设$y_1, y_2, \cdots, y_k$表示与待分类样本x距离最近的k个样本，假设样本的类别共有两类，那么$y_1, y_2, \cdots, y_k$中属于哪个类别的样本多，则将待分类样本x预测为哪个类别。

二、决策树

决策树（decision tree）是一种多级分类器，它可以将复杂的多类别分类问题转化为若干个简单的分类问题。它是一种树状结构，在每一个非叶子节点选取一个属性对样本进行分割，每一个叶子节点代表一个类别标签，如果一个样本落入一个叶子节点，则表明该样本属

于该叶子节点所代表的类别。

图2–28显示的是一个决策树分类器。在根节点中一共有295个乳腺癌患者样本，在非叶子节点通过E2F和KIAA0191这两个基因表达水平的高低，将295个乳腺癌患者分成"LOW RISK"、"MED RISK"、"HIGH RISK"三类，预测乳腺癌患者的生存情况。

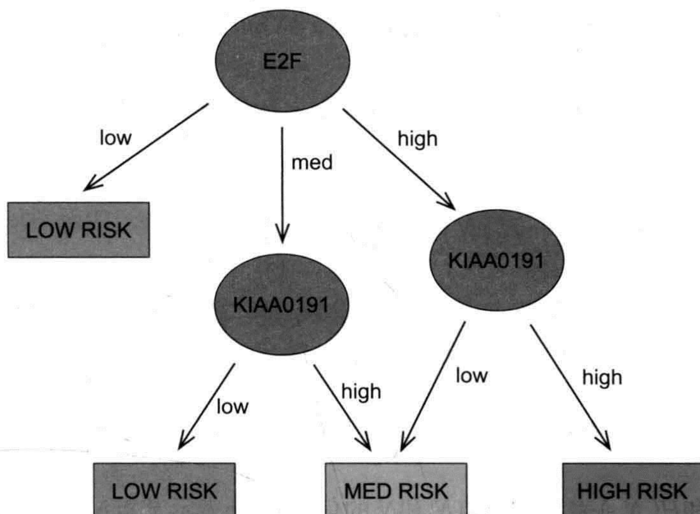


图2–28　决策树应用于乳腺癌基因表达谱的分类分析
来源于：Hallett RM，Hassell JA：E2F1 and KIAA0191 expression predicts breast cancer patient survival.BMC Res Notes.2011 Mar 31;4：95.

在决策树的构建过程中一般采用贪婪算法，自上而下地对样本进行递归分割。决策树的基本步骤如下：

1. 以代表所有训练样本的单个节点开始，如果样本属于同一类别，则该节点作为叶子节点。否则，依据某种分割规则选择最具分类能力的属性（如基因）作为决策树的当前节点。

2. 依据当前决策节点属性值（如基因表达水平）的不同，将训练样本分成若干子集。

3. 重复上面的步骤，使用递归的方法处理每个样本子集，直到符合终止条件为止。常用的终止条件包括所有叶子节点的样本都属于同一类别或叶子节点中包含了指定数目的样本（指定叶子节点应当包含的最少样本数）等。

利用基因芯片数据构造决策树的关键步骤在于每一个非叶子节点选取哪个基因以及用哪种分割规则对训练样本进行分类，这需要通过分割规则判断哪个基因更合适。分割规则主要包括：

*Gini*指数变化（$\Delta Gini$）：*Gini*指数是用来评价节点纯度的指标，将某节点*N*的*Gini*指数定义为：

$$Gini(N)=1-\sum_{i=1}^{k}p_i^2 \tag{2-17}$$

其中*pi*表示第*i*类样本在某节点中的概率，即某节点中第*i*类样本的频率；*k*表示样本类别的数量。*Gini*指数越小表示节点越纯，如果*Gini*指数为0则表示该节点中的所有样本属于同一个类别。

若节点N分割成两个字节点N_1和N_2，则$Gini$指数的变化值：

$$\Delta Gini = Gini(N) - \left(\frac{n_1}{n} Gini(N_1) + \frac{n_2}{n} Gini(N_2)\right) \tag{2-18}$$

其中$Gini(N_1)$和$Gini(N_2)$表示子节点N_1和N_2的$Gini$指数，n表示节点N中样本的个数，$n1$和$n2$分别表示N_1和N_2中样本的个数。通常选择$\Delta Gini$最大的基因作为分割属性以及对应的分割方式。

信息增益：该指标是用分割前后熵值的改变来评价节点纯度的变化。对于某节点N的信息熵定义为：

$$H(N) = -\sum_{i=1}^{k} p_i \log_2 p_i \tag{2-19}$$

其中pi表示第i类样本在某节点中的概率，k表示样本类别数。熵值越小说明节点越纯。

如果节点N分割成两个子节点N_1和N_2，则信息增益为：

$$Gain = H(N) - \left(\frac{n_1}{n} H(N_1) + \frac{n_2}{n} H(N_2)\right) \tag{2-20}$$

其中$H(N_1)$和$H(N_2)$表示子节点N_1和N_2的信息熵，n表示节点N中样本的个数，n_1和n_2分别表示N_1和N_2中样本的个数。通常选择信息增益最大的基因作为分割的属性以及对应的分割方式。

决策树分类器对训练样本集的准确率往往能够达到100%，但这会导致训练过度（对信号和噪声都适应），而且会让决策树生长的过于"枝繁叶茂"。既降低了决策树的可理解性和适用性，又使决策树本身对训练样本集过于依赖，一旦推广应用到新的数据时，决策树的准确性将迅速下降。因此限制决策树的生长和对决策树的修剪是极其必要的。常用的策略包括设定决策树的最大层数和设定每个节点包含的最小样本数等。决策树的修剪方法主要有前剪枝和后剪枝：前剪枝是在决策树的生成过程中通过设定阈值停止生长；后剪枝是在决策树长成以后由下而上进行修剪。

三、支持向量机

支持向量机（support vector machine，SVM）是由Vapnik等人在1995年提出的一种机器学习方法。它以统计学习理论为基础，根据结构风险最小化原则（structural risk minimization inductive principle，SRM）在选择的特征空间中构造最优超分面（optimal hyperplane），从而使未知样本的分类误差最小。

在很多情况下，训练样本集是线性不可分的，因此Vapnik等人提出了用高维分类面来解决这个问题。通过非线性变换将非线性问题转化为某个高维空间中的线性问题，在这个高维空间中寻找最优的分类面。而支持向量（support vector）对定义最优分类面极其重要，它们是过两类样本中离分类面最近的点、并且平行于最优分类面的超平面上的训练样本。在高维空间中分类函数只涉及训练样本之间的内积运算，而且这种内积运算可通过定义在原空间中的函数来实现，甚至不需要知道变换的形式。通过支持向量机得到的分类函数类似

于一个神经网络，其输出是一些中间层节点的线性组合，而每一个中间层节点对应于输入样本与一个支持向量的内积。最终的判别函数只包含与支持向量的内积和求和，因此识别的计算复杂度取决于支持向量的个数。

支持向量机通过选择的内积函数可实现线性和非线性分类。选择不同内积核函数将导致不同的支持向量机算法，目前比较常用的内积核函数主要有三类：

1. 多项式形式的内积核函数

$$K(x,x_i)=\left[(x\cdot x_i)+1\right]^q \tag{2-21}$$

此时获得的支持向量机是一个q阶多项式分类器。

2. 径向基内积核函数

$$K(x,x_i)=\exp\left\{-\frac{|x-x_i|^2}{\sigma^2}\right\} \tag{2-22}$$

得到一种径向基函数分类器。与传统的径向基函数方法的主要区别是，SVM中每个基函数的中心对应一个支持向量，它们以及输出权值都是由算法自行确定的。

3. S形函数内积核函数

$$K(x,x_i)=\tanh\left[v(x\cdot x_i)+c\right] \tag{2-23}$$

则支持向量机的实现形式是一个多层感知器神经网络，但是其中网络的权值、隐层节点数都是由算法自行确定的。

与传统的机器学习方法相比，支持向量机的主要优势有：

1. 支持向量机能够应用更多的距离（相似性）函数，其中包括线性函数和非线性函数来比较基因表达的测量值，从而能够更精确地考虑基因表达谱向量之间的关系。

2. 分类间隔的最大化，使得构建的分类模型具有较好的鲁棒性。

3. 支持向量机基于统计学习理论中结构风险最小化原理和VC维（Vapnik-Chervonenkis dimension）理论，具有较好的泛化能力，即通过有限的训练样本信息，在分类模型的复杂性和学习能力之间寻求最佳的折中，期望获得最优的推广能力。

四、分类器的分类效能评价

在分类的过程中，首先应用重抽样（re-sampling）技术把数据集分为训练集（training set）和测试集（test set）。利用训练集中的样本构建分类器，测试集用于评价通过训练集构建的分类器的分类效能。

（一）重抽样法

1. n倍交叉证实（n-fold cross validation） 将数据集随机分成近似相等的n份，选取其中的$n-1$份作为训练集构建分类器，剩下的一份作为测试集，如此循环n次。通过这种方法能够产生没有重复的训练集和测试集。

2. 留一法交叉证实(leave-one-out cross validation, LOOCV)　每次从数据集中随机抽取一个样本作为测试集,其余样本作为训练集。

3. Bootstrap aggregating　采取有放回抽样的方法,随机抽取不大于原数据集的样本集合(该集合成为原数据集的副本)。当随机抽样的数量和原数据集一致时,理论上每一个副本中包含原数据集63.2%的样本,剩余的为重复抽取的样本。将副本作为训练集,其余的样本作为测试集。

4. 无放回的随机抽样　每次随机抽取数据集的1/*n*作为测试集,其余样本作为训练集。

(二) 分类效能指标

1. 敏感性(sensitivity) $$\frac{TP}{TP+FN} \tag{2-24}$$

2. 特异性(specificity) $$\frac{TN}{TN+FP} \tag{2-25}$$

3. 阳性预测率(positive predictive value, precision) $$\frac{TP}{TP+FP} \tag{2-26}$$

4. 阴性预测率(negative predictive value) $$\frac{TN}{TN+FN} \tag{2-27}$$

5. 均衡正确率(balanced accuracy) $$\frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right) \tag{2-28}$$

6. 正确率(accuracy) $$\frac{TP+TN}{TP+TN+FP+FN} \tag{2-29}$$

其中*TP*(true positive)表示真阳性,即样本类别为阳性,分类器正确地将其判断为阳性的样本数; *TN*(true negative)表示真阴性,即样本类别为阴性,分类器正确地将其判断为阴性的样本数; *FP*(false positive)表示假阳性,即样本类别为阴性,分类器却错误地将其判断为阳性的样本数; *FN*(false negative)表示假阴性,即样本类别为阳性,分类器却错误地将其判断为阴性的样本数。

总之,对基因芯片数据进行分类分析有助于与疾病的精确诊断和预后分析。但是复杂疾病的发生往往不是由于单个基因的改变造成的,而是遗传因素和环境因素等共同作用产生的结果; 不同疾病涉及的基因不同,同种疾病在分子层面也存在着较大的异质性,因此,针对疾病相关的芯片数据进行分类分析时,如何选取合适的基因作为特征构建稳定的分类器以达到临床诊断的要求仍然是个极大的挑战。

第六节

基因芯片数据库及常用分析软件

Section 6 Microarray Databases and Softwares

一、基因表达数据库（gene expression omnibus，GEO）

近年来高通量检测基因表达的技术越来越成熟、应用也越来越广泛，例如，基因芯片，基因表达系列分析（serial analysis of gene expression，SAGE）和新一代测序（next generation sequences，NGS）等技术都可以实现对数以万计的基因转录本的检测。GEO是由美国国立生物技术信息中心（national center for biotechnology information，NCBI）开发和维护的公共数据库，它存储基因芯片数据、新一代测序数据以及其他形式的高通量功能基因组数据，并将其发布供研究者自由使用。目前，GEO储存了约20 000项研究得到的涉及500 000个样本、1300个物种、330亿单个基因的表达检测数据，这些数据是由世界各地的8000多个实验室提供的。基于web工具，用户可以对GEO存储的大量数据进行浏览、查询和可视化。通过四种编号GPL、GDS、GSE和GSM可以获得完整的平台、数据集、系列和样本的信息。例如，在Query部分常用GSE号输入到“GEO accession”中，可以了解芯片数据的详细信息（图2-29）。

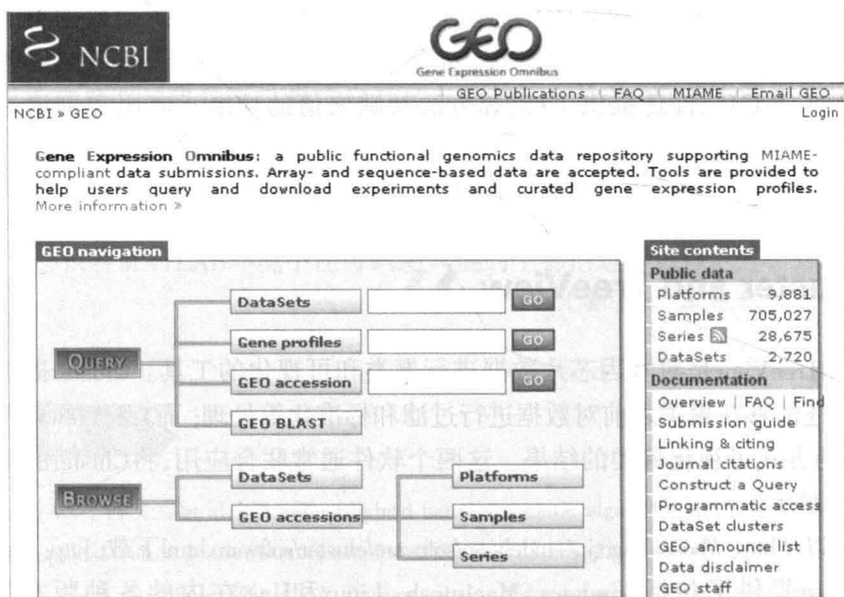


图2-29 GEO accession查询界面

GEO根据平台、数据集、系列和样本四种形式组织数据，使用数据的研究人员可以获取以下四方面相应的信息。

平台(platform, GPLxxx)：平台信息是由微阵列的简要描述和用来确定微阵列模板的数据表构成。最基本的平台信息是探针列表，它们规定了哪些基因可以在该芯片平台上被检测出来；平台编号以GPL为起始。

数据集(dataset, GDSxxx)：一个数据集中的样本来自相同的芯片平台，并且这些样本的检测值是都是以同种方式处理(如，背景校正和标准化等)之后得到的。数据集是由生物学上和统计学上能相互比较的样本所组成的，这些样本可能来自不同的数据提供者，它构成了GEO特有的数据显示和数据分析的基础；数据集编号以GDS为起始。

系列(series, GSExxx)：系列是由数据提供者提交给GEO的一次实验的基因芯片数据，这些数据具有明确的研究目的，是用户在使用GEO时经常采用的一种数据查询和下载方式；系列编号以GSE为起始。

样本(sample, GSMxxx)：在基因芯片实验中，一个样本中所有基因的表达水平通常由一张芯片来检测，样本信息由所检测的生物材料的描述、所遵循的实验协议和包含检测丰度值的数据表构成；样本编号以GSM为起始。

二、基因芯片显著性分析(significance analysis of microarray, SAM)

SAM是由美国Standford大学开发的一个免费软件，是目前使用最为广泛的差异表达基因筛选方法之一。SAM 软件以插件的形式在Excel中运行，使用简单、很容易被生物医学工作者所掌握。SAM考虑到基因芯片数据噪声大小与表达丰富相关的特点，对t检验进行修正，为每个基因计算一个统计量D，其表示该基因表达水平均值的变化(比如，在疾病正常两类之间的变化)与标准差的比值。此外，该方法使用随机扰动数据集的方法估计随机情况下统计量D的分布，通过选择*delta*值，确定FDR的水平，从而识别显著差异表达的基因。此外，SAM软件还提供了k近邻方法补缺失值的功能。应用实例请参见本章第三节。

SAM可以从http://www-stat.stanford.edu/~tibs/SAM/index.html下载。

三、Cluster and TreeView

Cluster和TreeView是对基因芯片数据进行聚类和可视化的工具。Cluster提供了多种聚类算法，同时还能够在聚类之前对数据进行过滤和标准化等处理；而TreeView则能够以热图和系统树图的方式可视化聚类的结果。这两个软件通常联合应用，将Cluster的聚类结果用TreeView进行显示。

Cluster可以从http://bonsai.hgc.jp/~mdehoon/software/cluster/software.html下载；http://www.treeview.net/tv/download.asp提供了包括Windows, Macintosh, Linux和Unix在内的各种版本的TreeView软件。

四、BRB-ArrayTools

BRB-ArrayTools是一款综合的基因芯片数据分析软件。BRB-ArrayTools能够针对多种平台的基因表达谱数据进行几乎所有的常规数据分析，包括预处理、标准化、聚类、分类、功能注释、可视化等。BRB-ArrayTools也是以Excel加载宏的形式呈现，所以操作简单、使用方面。上面提到的SAM、Cluster和TreeView均已整合到ArrayTools软件中。

ArrayTools可以从http: //linus.nci.nih.gov/BRB-ArrayTools.html下载。

五、R语言和Bioconductor

R语言是一种计算机程序设计语言，也是一个开放式的软件开发平台。R语言具有强大的数学统计分析和科学数据可视化功能，能提供各种数据处理、统计分析及图形显示工具。R语言在生物信息领域具有重要的应用价值，利用R语言可以进行基因芯片数据的差异表达分析、聚类分析和分类分析等。软件研究人员可以在R语言这个开放平台上不断扩充其功能，开发出面向特定应用的软件。

Bioconductor是一个基于R语言的、面向基因组信息分析的应用软件集合。Bioconductor的应用功能是以包的集成形式呈现给用户，它提供的软件包中包括各种基因组数据分析和注释工具。同时，Bioconductor还提供了许多专门的基因芯片分析软件包，可以实现数据的预处理、各种分析、注释及可视化等功能。Affy是分析Affymetrix寡聚核苷酸芯片的软件包，可用于数据的读取、过滤、标准化等。Marray是用于双通道（cDNA）微阵列数据的预处理软件包。Limma包通过使用线性模型来分析设计实验和评估差异表达，可应用于所有类型的芯片数据。

六、Matlab: Bioinformatics Toolbox

Matlab是美国MathWorks公司出品的商业数学软件，是用于算法开发、数据可视化、数据分析以及数值计算的高级技术计算语言和交互式环境。Matlab的基本数据单位是矩阵，而基因表达谱也是矩阵形式的数据，因此，通过Matlab编程能够比较容易地进行基因芯片的数据分析。其中，Bioinformatics Toolbox是基于MATLAB环境开发的基因组和蛋白质组分析工具箱。该工具箱功能强大，可进行数据库访问、序列比对、基因芯片数据分析、可视化以及功能注释等。此外，在MATLAB环境中还可调用其他的生物信息学软件。

（姜 伟 王艳秋 陈晓文 杨 磊）

参考文献

1. Li, L. , Jiang W. , Li X. , et al. , A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. Genomics, 2005. 85(1): 16-23.

2. Leping Li, Clarice R. Weinberg, Thomas A. Darden, et al. , *Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method.* Bioinformatics, 2001.

17(12): 1131–1142.

3. Isabelle Guyon, Jason Weston, Stephen Barnhill, M. D. and Vladimir Vapnik. . *Gene Selection for Cancer Classification using Support Vector Machines.* Machine Learning, 2002. 46(1): 389–422.
4. Li X, Rao S, Wang Y, Gong B. . *Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling.* Nucleic Acids Res, 2004. 32(9): 2685–94.
5. Zeng, Q, E. Burdet, C. L. Teo, *Evaluation of a collaborative wheelchair system in cerebral palsy and traumatic brain injury users.* Neurorehabil Neural Repair, 2009. 23(5): 494–504.
6. De Hoon, M. J. *Open source clustering software.* Bioinformatics, 2004. 20(9): 1453–4.
7. Hallett, R. M. and J. A. Hassell, *E2F1 and KIAA0191 expression predicts breast cancer patient survival.* BMC Res Notes, 2011. 4 : 95.

第三章

CHAPTER 3

基因注释与功能分类

GENE ANNOTATION AND FUNCTIONAL CLASSIFICATION

第一节 引 言

Section 1 Introduction

随着后基因组(post-genomics)时代的来临,基因组学的研究重心开始从阐明所有遗传信息转移到在整体分子水平对功能进行研究。这种转变的一个重要标志是产生了功能基因组学(functional genomics)。功能基因组学利用结构基因组所提供的信息和产物,发展和应用新的实验手段,通过在基因组或系统水平上全面分析基因的功能,使得生物学研究从对单一基因或蛋白质的研究转向多个基因或蛋白质同时进行系统的研究。功能基因组学的主要任务之一是进行基因组功能注释(genome annotation),了解基因的功能,认识基因与疾病的关系,掌握基因的产物及其在生命活动中的作用等。在使用全局方法进行研究时,研究人员往往同时检测大量基因的表达水平,从而在整体水平上获得关于基因功能及基因之间相互作用的信息,如何应用生物信息学方法,高通量地注释这些基因的生物学功能是一个重要的挑战。快速有效的基因注释对进一步识别基因,识别基因转录调控信息,研究基因的表达调控机制,研究基因在生物体代谢途径中的地位,分析基因、基因产物之间的相互作用关系,绘制基因调控网络图,预测和发现蛋白质功能,揭示生命的起源和进化等具有重要的意义。

本章主要介绍当前常用的基因注释数据库和基因功能预测方法,以及在此基础上发展起来的基因集功能富集分析、基因功能比较等方法和常用工具。

第二节

基因注释数据库

Section 2 Gene Annotation Database

一、GO(gene ontology)数据库

随着生物技术和信息技术的发展,研究人员已经掌握了大量的全基因组数据,同时关于基因、基因产物以及基因功能知识越来越丰富,这些知识被生物学家共享,如何利用这些先验知识,使之成为计算机可识别并操作的资源,这需要合理组织和系统的方法。因此提供一个结构化的标准的生物学模型,以便计算机程序进行分析,成为从整体水平系统研究基因及其产物的一项基本需求。本节主要介绍当前应用较为广泛的基因及其产物注释数据库: GO。

GO目标是建立一个可以适用于各种物种的,对基因和蛋白质功能进行限定和描述的,动态控制的词表,即使关于某个基因或蛋白的功能与作用的知识未知或在不断积累变化中,我们仍然能有一定的规则去描述更新它。GO中有约束的功能词汇(terms)称为一个概念,它表示一个功能类,使用它来描述众多的基因的功能,并严格地定义功能类之间的关系。功能类之间的关系分为is-a,和part-of两类, is-a表示子功能是父功能的一个实例, part-of表示子功能是父功能的一个部分。GO以有向无环图方式表示功能类之间的关系(图3-1),它的

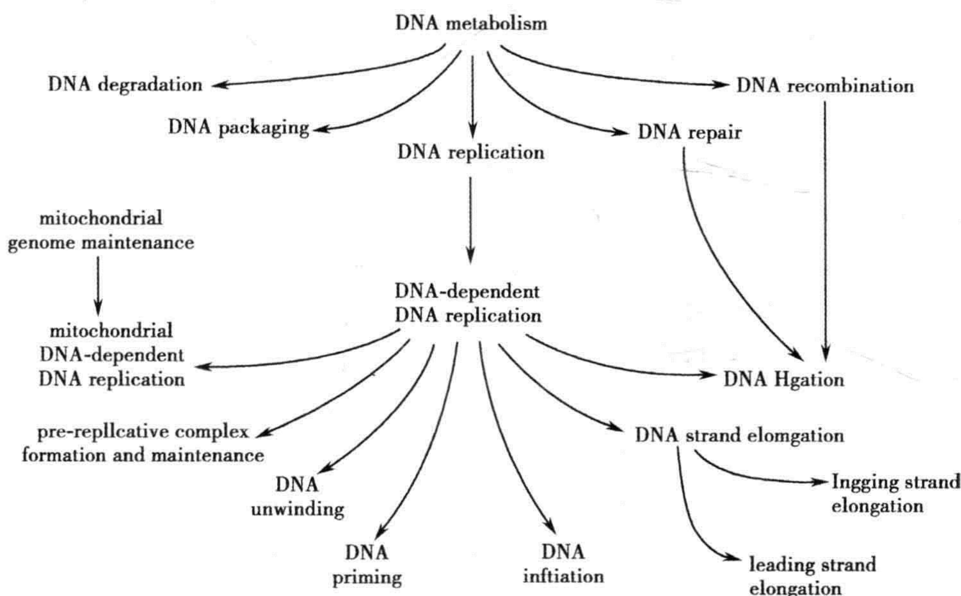


图3-1 GO中生物学过程的DNA代谢部分功能类示意图

一个子结点可以有多个父结点，但没有循环关系，父辈结点包含所有子结点的含义，即从父结点到子结点，含义是逐层深入的关系。

（一）GO数据库构成

GO将基因功能划分为细胞组分（cellular component）、分子功能（molecular function）、生物学过程（biological process）3个分支（表3-1）。因此，一个基因或蛋白可从三个层面得到注释，可能是同一个物体存在的多种性质。如细胞色素c，在分子功能上体现为电子传递活性，在生物过程中与氧化磷酸化和细胞凋亡有关，在细胞中存在于线粒体质中和线粒体内膜上。随着生命科学研究的逐步深入，GO注释数据库正在不断积累和更新。目前GO已经成为生物信息领域中一个重要的方法和工具，并正在逐步改变着我们对各种生物学数据的组织和理解方式，它的存在已经大大加快了生物数据的整合和利用。

项目最初是由1988年对三个模式生物数据库的整合开始：果蝇数据库（FlyBase）、酵母基因组数据库（saccharomyces genome database）和小鼠基因组数据库MGD（the mouse genome database），随后相继收录了更多数据，GO不断发展扩大，现在已包含数十个动物、植物、微生物的数据库。GO术语在多个合作数据库中的统一使用，促进了各类数据库对基因描述的一致性。目前已经成为应用最广泛的基因注释体系之一。

表3-1 GO数据库收录的基因组数据列表

机构简称	收录的基因组数据	网站
BBOP	果蝇	http: //www.berkeleybop.org
BHF-UCL	心血管基因	http: //www.cardiovasculargeneontology.com
dictyBase	粘菌盘基网柄菌	http: //dictybase.org
EcoliWiki	大肠埃希菌	http: //ecoliwiki.net
FlyBase	果蝇	http: //flybase.bio.indiana.edu
GeneDB	裂殖酵母、恶性疟原虫、硕大利什曼原虫、布氏锥虫	http: //www.genedb.org
GOA	UniProt和InterPro注释	http: //www.ebi.ac.uk/GOA
Gramene	农作物基因数据库	http: //www.gramene.org
MGD and GXD	小家鼠	http: //www.informatics.jax.org
RGD	褐家鼠	http: //rgd.mcw.edu
Reactome	生物过程知识库	http: //www.genomeknowledge.org
SGD	芽殖酵母、酿酒酵母	http: //www.yeastgenome.org
TAIR	拟南芥	http: //www.arabidopsis.org
IGS	基因组研究的工具和数据	http: //www.igs.umaryland.edu
JCVI	若干种细菌基因组数据库	http: //www.jcvi.org
WormBase	线虫	http: //www.wormbase.org
ZFIN	斑马鱼	http: //zfin.org

(二) GO数据库的在线注释

GO中的结点如何与对应的基因产物相联系呢？这是由参与合作的数据库来完成的，它们使用GO的定义方法，对它们所包含的基因产物进行注释，并且提供支持这种注释的参考和证据。每个基因或基因产物都会有一个列表，列出与之相关的GO结点。现在对于基因或者结点的注释可以使用多种不同的工具软件进行查询，它们大多数GO浏览器都是web模式的，允许你直观地看到结点和其相关信息，如定义、同义词和数据库参考等。有些GO浏览器如AlliGO和QuickGo，可以看到每个结点的注释。

我们这里使用AmiGO作为实例说明GO数据库的在线注释。在GO数据库中，每条记录都有一个数据标识号GO: XXXXXX和对应的结点。因此检索时需要知道待查基因的名字或结点的数字标识号，将它们直接输入检索框即可。如果检索的基因或蛋白质存在别名，可在检索框下勾"gene or proteins"，并在检索框中输入别名检索；"exact match"表示是否完全匹配，可供选择。

这里以检索神经源性分化因子6(NEUROD6)为例。在检索框中输入"NEUROD6"并勾选"gene and proteins"和"exact match"，运行后所得基因产物。检索得到的四个记录分别是不同物种中的神经源性分化因子6，点击物种为人类的"NEUROD6"记录，即为该基因产物的基本信息，包括类型、物种、别名来源和序列；图3-2显示了该基因产物的结点关联(term associations)图，图中记录名称"Term"是GO记录的名字，"Ontology"是该基因产物的特性，如要查看其分子功能，可点击其中的一条记录"nervous system development"（图3-3）。

Term Associations

Download all association information in: gene association format RDF-XML

▼ Filter associations displayed

Filter Associations — Ontology: All / biological process / cellular component / molecular function; Evidence Code: All / IC / IDA / IEA

Set filters | Remove all filters

Select all | Clear all | Perform an action with this page's selected terms | Go!

Accession, Term	Ontology	Qualifier	Evidence	Reference	Assigned by
9965 gene products view in tree GO:0030154 : cell differentiation	biological process		IEA With SP KW:KW-0221	GO REF:0000004	UniProtKB (via UniProtKB/Swiss-Prot)
23925 gene products view in tree GO:0007275 : multicellular organismal development	biological process		IEA With SP KW:KW-0217	GO REF:0000004	UniProtKB (via UniProtKB/Swiss-Prot)
5317 gene products view in tree GO:0007399 : nervous system development	biological process		IEA With SP KW:KW-0524	GO REF:0000004	UniProtKB (via UniProtKB/Swiss-Prot)
20473 gene products view in tree GO:0045449 : regulation of cellular transcription	biological process		IEA With SP KW:KW-0805	GO REF:0000004	UniProtKB (via UniProtKB/Swiss-Prot)
36436 gene products view in tree GO:0005634 : nucleus	cellular component		IEA With SP SL:SL-0191	GO REF:0000023	UniProtKB (via UniProtKB/Swiss-Prot)
21250 gene products view in tree GO:0003677 : DNA binding	molecular function		IEA With SP KW:KW-0238	GO REF:0000004	UniProtKB (via UniProtKB/Swiss-Prot)
16342 gene products view in tree GO:0030528 : transcription regulator activity	molecular function		IEA With InterPro:IPR011598	GO REF:0000002	UniProtKB (via UniProtKB/Swiss-Prot)

Select all | Clear all | Perform an action with this page's selected terms | Go!

Back to top

图3-2 AmiGO基因描述示例

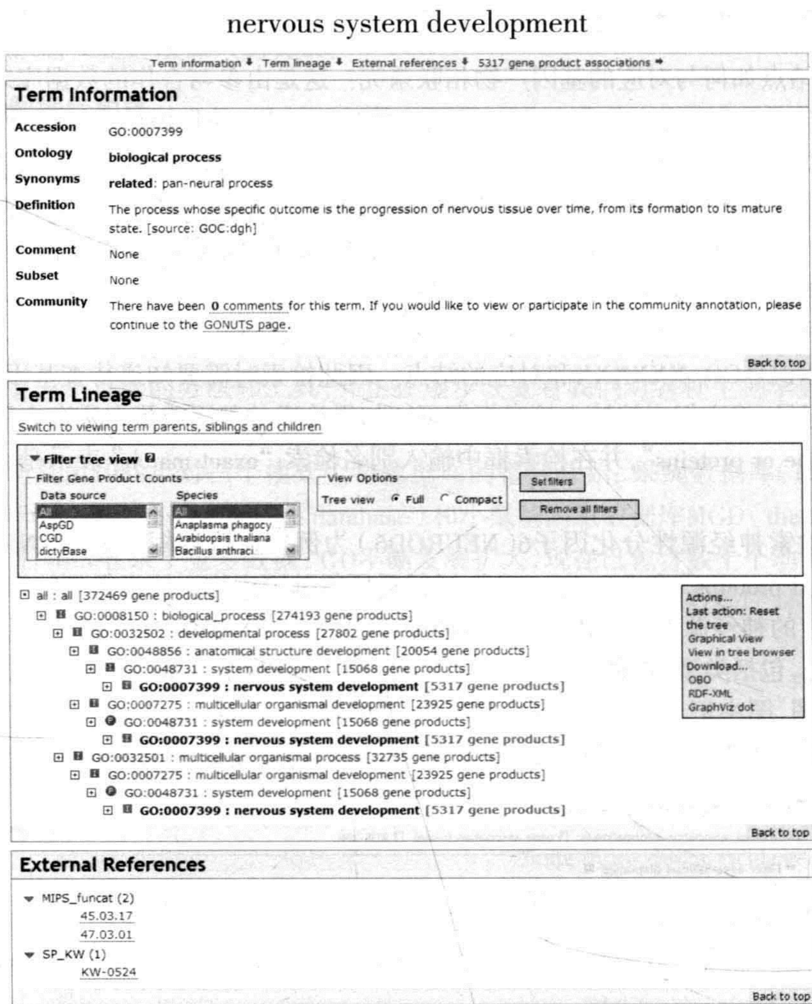


图3-3 AmiGO基因功能描述示例

图3-3上部先对神经源性分化因子6的相关信息做简单描述，中间结点系谱（Term Lineage）呈阶梯状分布，记录了GO数据库中全部分子功能所处的位置和关系。下方“External Reference”提供了与外部相关数据的链接。点击右上方的可视化视图（Graphical View）就更清晰地显示了分子功能记录之间构成的复杂网状结构，既有上下隶属关系，也存在平行关系（图3-4）。

对于未知基因名的序列，可以用序列直接检索GO 数据库。点击AmiGO首页上方的“BLAST”，在检索框中输入氨基酸或核酸序列，网页能自动识别并相应地做BLASTP或BLASTX和数据库中的序列比对，比对到序列相似的基因，同上面的做法一样，可以查询到功能注释信息。

（三）GO数据库本地化及批量注释

GO的所有数据都是免费获得的。GO数据中包含了结点间的结构（ontologies）和基因或基因产物的注释（annotations）数据，还包括蛋白序列比对的数据（图3-5）。其中结构数据包含结点和结点之间的连接关系，注释数据包含由数据库成员提交的基因或基因产物与结点

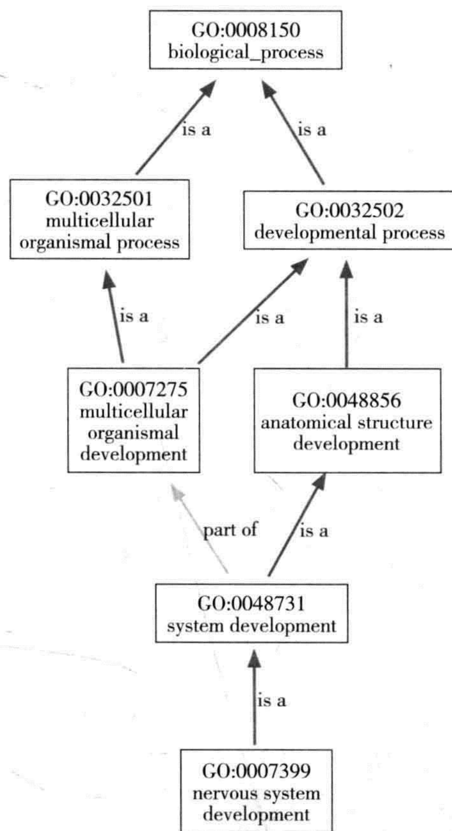


图3-4 AmiGO查询结果图形视图

Build name	Contents	Updated
termdb	ontologies, definitions and mappings to other dbs	daily
assocdb	termdb (above); all manual gene product annotations; electronic annotations (IEA) from all databases **other than** UniProtKB	weekly
seqdb	assocdb (above), plus protein sequences for most of the gene products	weekly
full GO database	termdb (above), plus manual **and** electronically generated (IEA) annotations	monthly

图3-5 可供下载的GO数据

间关联。结构和注释两种数据分别储存在单独的数据库中，这使得利用结构对注释的查询更加有效。

GO的网站上提供多种数据形式的下载：MySQL，OBO XML，OWL，RDF XML，SQL。XML和MySQL文件是被储存于独立的GO数据库中（图3-6）。下载数据到本地后，如果需要找到与某一个GO术语相关的基因或基因产物，可以找到一个相应表格，搜寻到这种注解的编号，并且可以链接到与之对应的位于不同数据库的基因相关文件。

当用户希望对大量的基因进行基因注释时，GO的网站上提供了许多推荐的工具，可以基于GO做批量分析。我们以GENETOOLS为例，它可以提供基因和结点的批量查询，用以方便用户对GO注释的解释，它还提供了注释结点树状结构的可视化，并能自由编辑（图3-7）。

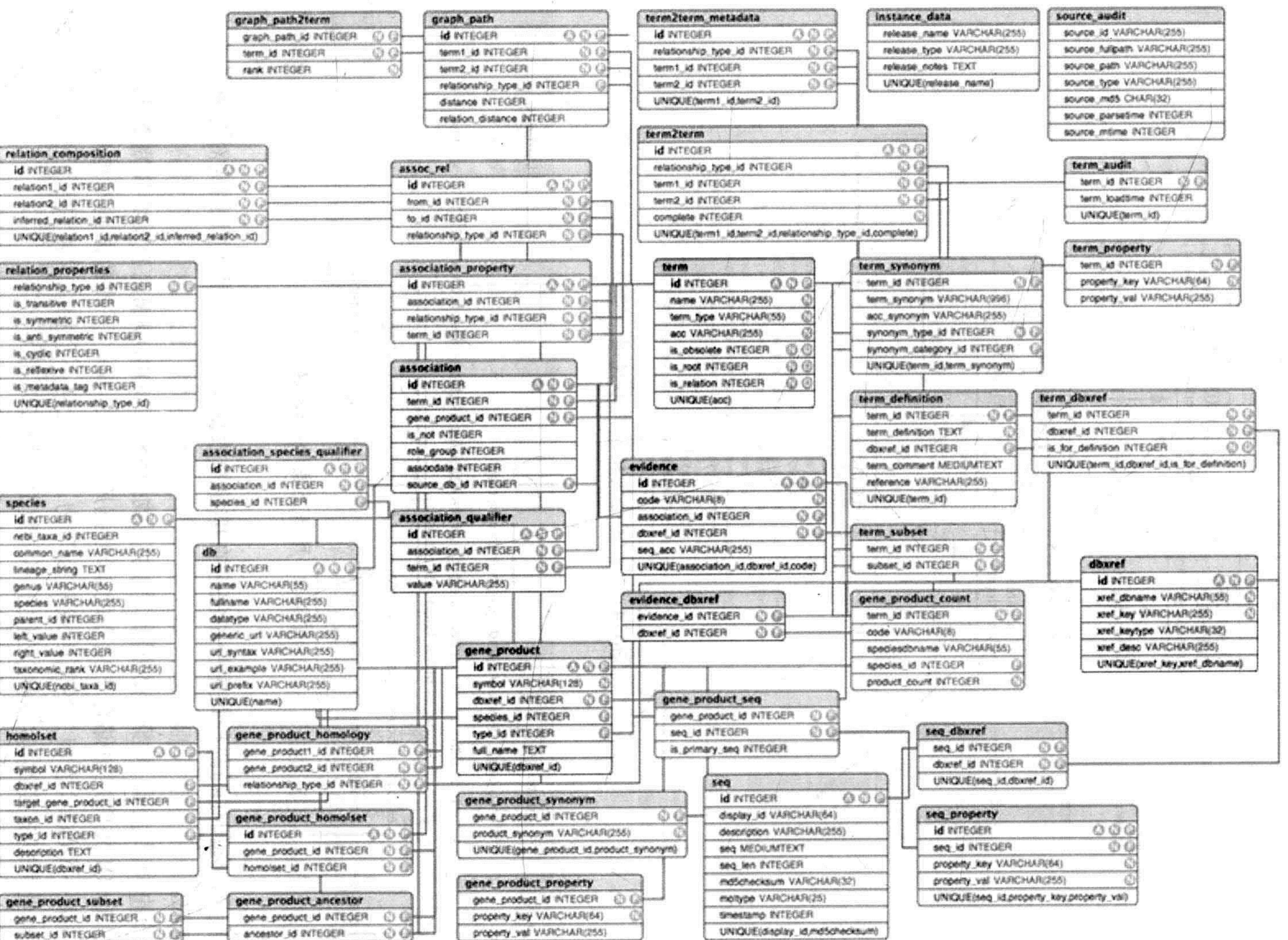

graph_path2term
graph_path_id INTEGER
term_id INTEGER
rank INTEGER
graph_path
id INTEGER
term1_id INTEGER
term2_id INTEGER
relationship_type_id INTEGER
distance INTEGER
relation_distance INTEGER
term2term_metadata
id INTEGER
relationship_type_id INTEGER
term1_id INTEGER
term2_id INTEGER
UNIQUE(term1_id,term2_id)
instance_data
release_name VARCHAR(255)
release_type VARCHAR(255)
release_notes TEXT
UNIQUE(release_name)
source_audit
source_id VARCHAR(255)
source_fullpath VARCHAR(255)
source_path VARCHAR(255)
source_type VARCHAR(255)
source_md5 CHAR(32)
source_parsetime INTEGER
source_mtime INTEGER
relation_composition
id INTEGER
relation1_id INTEGER
relation2_id INTEGER
inferred_relation_id INTEGER
UNIQUE(relation1_id,relation2_id,inferred_relation_id)
assoc_rel
id INTEGER
from_id INTEGER
to_id INTEGER
relationship_type_id INTEGER
term2term
id INTEGER
relationship_type_id INTEGER
term1_id INTEGER
term2_id INTEGER
complete INTEGER
UNIQUE(term1_id,term2_id,relationship_type_id,complete)
term_audit
term_id INTEGER
term_loadtime INTEGER
UNIQUE(term_id)
term_property
term_id INTEGER
property_key VARCHAR(64)
property_val VARCHAR(255)
relation_properties
relationship_type_id INTEGER
is_transitive INTEGER
is_symmetric INTEGER
is_anti_symmetric INTEGER
is_cyclic INTEGER
is_reflexive INTEGER
is_metadata_tag INTEGER
UNIQUE(relationship_type_id)
association_property
id INTEGER
association_id INTEGER
relationship_type_id INTEGER
term_id INTEGER
term
id INTEGER
name VARCHAR(255)
term_type VARCHAR(55)
acc VARCHAR(255)
is_obsolete INTEGER
is_root INTEGER
is_relation INTEGER
UNIQUE(acc)
term_synonym
term_id INTEGER
term_synonym VARCHAR(996)
acc_synonym VARCHAR(255)
synonym_type_id INTEGER
synonym_category_id INTEGER
UNIQUE(term_id,term_synonym)
association
id INTEGER
term_id INTEGER
gene_product_id INTEGER
is_not INTEGER
role_group INTEGER
assocdate INTEGER
source_db_id INTEGER
term_definition
term_id INTEGER
term_definition TEXT
dbxref_id INTEGER
term_comment MEDIUMTEXT
reference VARCHAR(255)
UNIQUE(term_id)
term_dbxref
term_id INTEGER
dbxref_id INTEGER
is_for_definition INTEGER
UNIQUE(term_id,dbxref_id,is_for_definition)
association_species_qualifier
id INTEGER
association_id INTEGER
species_id INTEGER
evidence
id INTEGER
code VARCHAR(8)
association_id INTEGER
dbxref_id INTEGER
seq_acc VARCHAR(255)
UNIQUE(association_id,dbxref_id,code)
species
id INTEGER
ncbi_taxa_id INTEGER
common_name VARCHAR(255)
lineage_string TEXT
genus VARCHAR(55)
species VARCHAR(255)
parent_id INTEGER
left_value INTEGER
right_value INTEGER
taxonomic_rank VARCHAR(255)
UNIQUE(ncbi_taxa_id)
association_qualifier
id INTEGER
association_id INTEGER
term_id INTEGER
value VARCHAR(255)
term_subset
term_id INTEGER
subset_id INTEGER
db
id INTEGER
name VARCHAR(55)
fullname VARCHAR(255)
datatype VARCHAR(255)
generic_url VARCHAR(255)
url_syntax VARCHAR(255)
url_example VARCHAR(255)
url_prefix VARCHAR(255)
UNIQUE(name)
dbxref
id INTEGER
xref_dbname VARCHAR(55)
xref_key VARCHAR(255)
xref_keytype VARCHAR(32)
xref_desc VARCHAR(255)
UNIQUE(xref_key,xref_dbname)
evidence_dbxref
evidence_id INTEGER
dbxref_id INTEGER
gene_product_count
term_id INTEGER
code VARCHAR(8)
speciesdbname VARCHAR(55)
species_id INTEGER
product_count INTEGER
gene_product
id INTEGER
symbol VARCHAR(128)
dbxref_id INTEGER
species_id INTEGER
type_id INTEGER
full_name TEXT
UNIQUE(dbxref_id)
gene_product_seq
gene_product_id INTEGER
seq_id INTEGER
is_primary_seq INTEGER
seq_dbxref
seq_id INTEGER
dbxref_id INTEGER
UNIQUE(seq_id,dbxref_id)
homolset
id INTEGER
symbol VARCHAR(128)
dbxref_id INTEGER
target_gene_product_id INTEGER
taxon_id INTEGER
type_id INTEGER
description TEXT
UNIQUE(dbxref_id)
gene_product_homology
gene_product1_id INTEGER
gene_product2_id INTEGER
relationship_type_id INTEGER
seq
id INTEGER
display_id VARCHAR(64)
description VARCHAR(255)
seq MEDIUMTEXT
seq_len INTEGER
md5checksum VARCHAR(32)
moltype VARCHAR(25)
timestamp INTEGER
UNIQUE(display_id,md5checksum)
gene_product_synonym
gene_product_id INTEGER
product_synonym VARCHAR(255)
UNIQUE(gene_product_id,product_synonym)
gene_product_homolset
id INTEGER
gene_product_id INTEGER
homolset_id INTEGER
seq_property
id INTEGER
seq_id INTEGER
property_key VARCHAR(64)
property_val VARCHAR(255)
UNIQUE(seq_id,property_key,property_val)
gene_product_subset
gene_product_id INTEGER
subset_id INTEGER
gene_product_ancestor
gene_product_id INTEGER
ancestor_id INTEGER
gene_product_property
gene_product_id INTEGER
property_key VARCHAR(64)
property_val VARCHAR(255)

图3-6 GO数据结构图

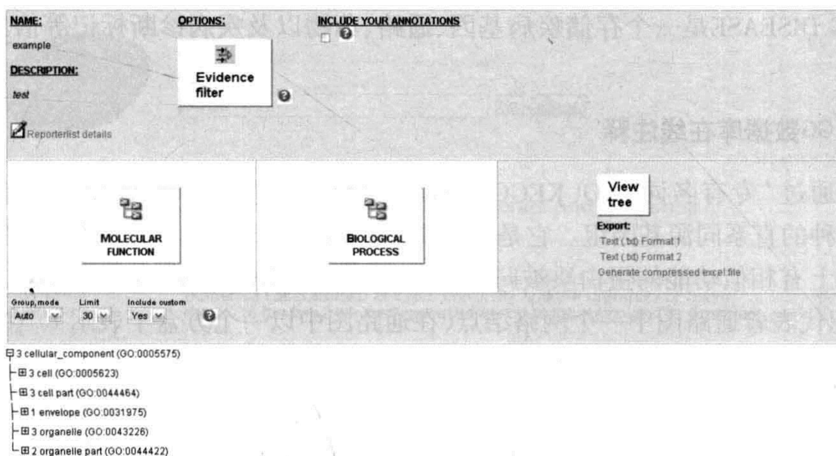


图3-7 GENETOOLS工具批量查找示例

二、KEGG数据库

生物体细胞的生物学功能是许多分子相互作用的结果，不能仅仅归功于单个基因或单个分子。KEGG（Kyoto encyclopedia of genes and genomes，京都基因与基因组百科全书）就是将基因组中的一系列基因用一个细胞内的分子相互作用的网络连接起来的过程，如一个通路或是一个复合物，通过它们来展现更高一级的生物学功能的数据库。KEGG将基因组信息和高一级的功能信息有机地结合起来，通过对细胞内已知生物学过程的计算机化处理和将现有的基因功能解释标准化，整合了基因组学、生物化学以及系统功能组学的信息，有助于研究者把基因及表达信息作为一个整体网络进行研究。

（一）KEGG数据库的主要组成

KEGG中的pathway是根据相关知识手绘的，这里的手绘的意思可能是指人工以特定的语言格式来确定通路各组件的联系；基因组信息主要是从NCBI等数据库中得到的，除了有完整的基因序列外，还有没完成的草图。KEGG目前共包含了19个子数据库，它们被分类成系统信息、基因组信息和化学信息三个类别。①基因组信息存储在GENES数据库里，包括全部完整的基因组序列和部分测序的基因组序列，并伴有实时更新的基因相关功能的注释，更高级的功能信息则存储在PATHWAY数据库里，包括图解的细胞生化过程如代谢、膜转运、信号传递、细胞周期和同系保守的子通路等信息；一些直系同源的基因数据作为PATHWAY数据库的补充，形成了PATHWAY数据库中一些保守的子通路（pathway motifs），这些子通路通常有一些在染色体位置上邻近的基因编码，这对于基因功能的预测十分重要；②KEGG中化学信息的6个数据库被称为KEGG LIGAND数据库，包含化学物质、酶分子、酶化反应等信息。KEGG BRITE数据库是一个包含多个生物学对象的基于功能进行等级划分的本体论数据库，它包括分子、细胞、物种、疾病、药物以及它们之间的关系，该数据库将基因与外界环境影响联系起来。例如，可以通过BRITE数据库分析药物和靶点之间的关系。③一些小的通路模块被存储在MODULE数据库中，该数据库还存储了其他的一些相关功能的模块以及化合物信息；④KEGG DRUG数据库存储了目前在日本所有非处方药和美国的大部分处方药

品; ⑤KEGG DISEASE是一个存储疾病基因、通路、药物以及疾病诊断标记等信息的新型数据库。

(二) KEGG数据库在线注释

KEGG 通过"专有名词" KO(KEGG orthology)对基因进行注释,每个KO标识代表一个来自不同物种的直系同源基因组。它是蛋白质(酶)的一个分类体系,序列高度相似,并且在同一条通路上有相似功能的蛋白质被归为一组,然后打上KO(或K)标签。在KEGG通路中,每个KO标识代表着通路图中一个网络结点(在通路图中以一个方盒子表示)。在KEGG对每个对象的功能及其他等级划分中,KO标识则代表着底层的叶子结点。

KO标识是基因组通过KEGG通路以及KEGG等级划分与生物学系统关联的基础。对于KEGG中的每个物种来说,物种特异性通路以及功能等级的划分是通过计算的方法自动实现的,在这一过程中KO标识是必不可少的。有了这些物种特异性通路以及功能等级划分,由基因芯片表达谱等高通量方法得到的基因便可以注释到相应的位置,以此来系统的分析该基因在细胞或组织中的功能。除了对基因或蛋白的功能等级划分之外, KEGG BRITE数据库还包含了化合物(C、D、G、R标识)以及其作用关系的等级划分。

KO标识还可以将基因的基因组信息以及转录组信息与通路总化合物分子的化学结构联系起来,因此, KO分类系统还可以应用化学信息注释上。这一过程实现的基本原理是每个KO下的基因所标识的酶是不同的,其对应化学底物也不同,另外,还有对生物合成通路信息的不断积累、不断更新作为数据支撑的基础。例如: 糖类的生物合成是通过一系列的生化反应来完成的,这些反应都是由糖基转移酶催化。在KEGG PATHWAY中,与糖类生物合成相关的通路图中各种糖类相关的化合物都是通过一条边与糖基转移酶的一组同源基因(KO group)直接相连,一旦在通路中确定了基因的注释位置,则与其相关的糖类化合物也被找到。应用相似的方法可以对基因芯片表达谱数据进行糖类结构以及功能的预测,这一方法已被广泛使用。除了糖类化合物之外,在KEGG数据库中还存储了很多其他化合物(多聚不饱和脂肪酸、萜类化合物、聚酮化合物等)的结构和功能信息,通过以上方法可以对基因进行化学信息的注释。

下面以人类亚甲基四氢叶酸还原酶(methylenetetrahydrofolate reductase, MTHFR)为例: 首先进入KEGG首页,在首页顶端的输入框中输入人类亚甲基四氢叶酸还原酶名称"MTHFR" (图3-8)。

点击搜索按钮"GO"进入查询结果页面(图3-9),该页面会列出针对基因"MTHFR"在KEGG数据库中的搜索结果,除人类外,包含"MTHFR"基因的物种条目也会被列出。

其中排在第一位的是人类基因"MTHFR"的相关信息,点击该条目进入到详细信息页面(图3-10)。

该页面以表格的形式列出了该基因有关的详细信息,包括基因编号,基因的详细定义,所编码酶的编号,基因所在通路,以及序列的编码信息。同时,在页面的右侧还提供了该基因在其他分子生物学数据库的链接,如OMIM、NCBI、GenBank等。

通过点击相应的链接,我们可以进入该基因相应信息的页面。在pathway这一栏中列出了该基因所在的生物学通路,点击编号为hsa00670 One carbon pool by folate通路,进入到该通路的相应页面(图3-11)。



图3-8　KEGG查询首页

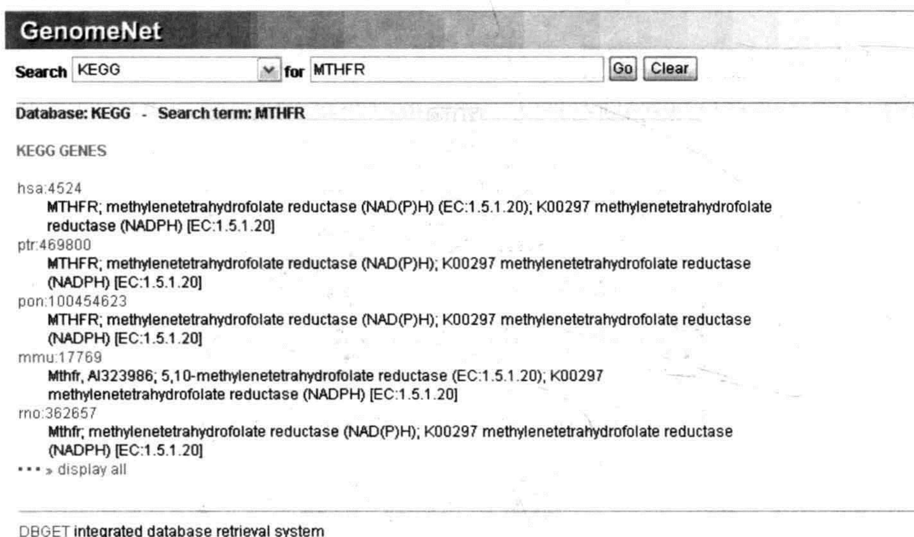


图3-9　MTHFR基因的KEGG通路查询结果

该编号为hsa00670的通路页面以简单的几何图形显示出相关生物过程。图中红色的方框即为基因"MTHFR"所编码的酶，方框里面的1.5.1.20是EC编号；小圆圈代表代谢物，鼠标放上会出现C×××××，C代表compound，五位数编号×××××是这种化合物在KEGG中的编号，大的圆方块，表示另一个代谢图，绿色的方框表示这个物种特有的基因或酶。以此就可以通过该酶所在位置以及通路的拓扑结构来综合分析基因。此外，可以通过页面顶部

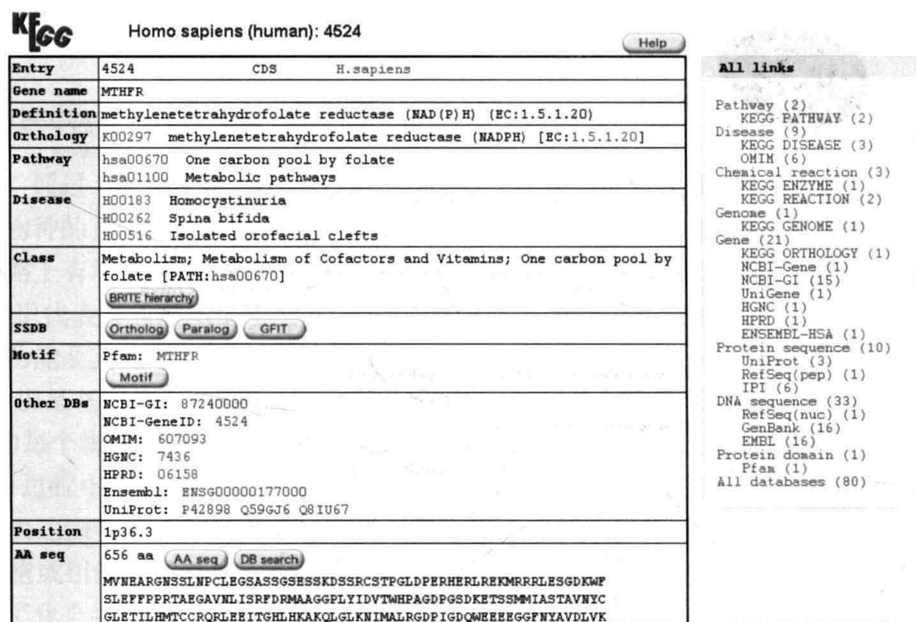


图3-10 人类基因“MTHFR”的详细信息

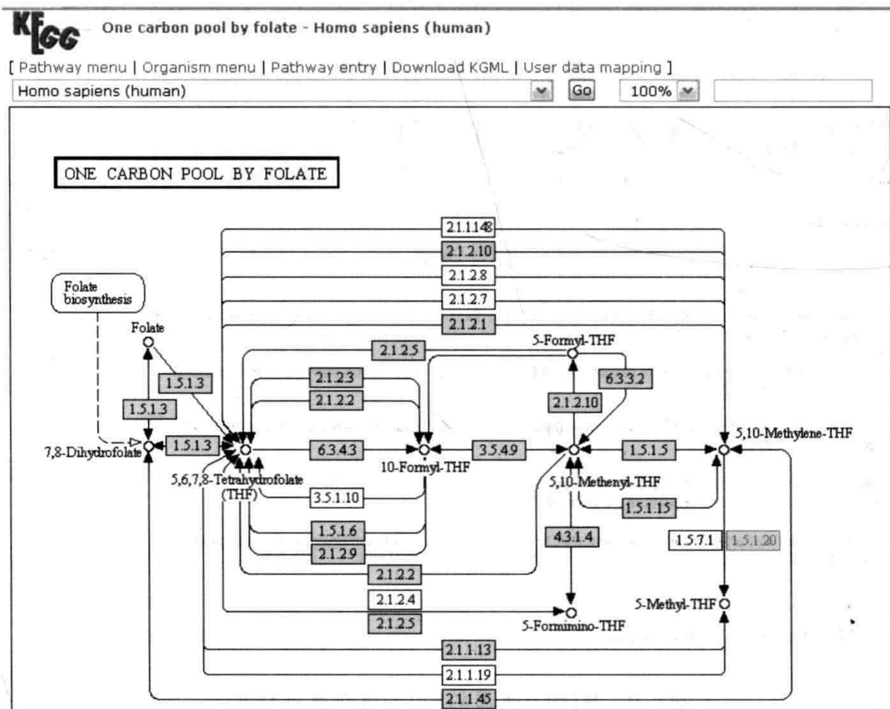


图3-11 人类One carbon pool by folate通路

的下拉列表框来选择该通路在其他物种中的信息，也可以通过该列表框的选择来查看相关的基因、酶、反应、化合物等相关通路信息。

点击通路图上方的pathway entry，在出现的页面中点击pathway map按钮链接Ortholog table，就进入了Ortholog table如下的页面（图3-12）：

Ortholog table(ko00670)

Eukaryotes Page: 1

Organism	K00287 (folA) [105]	K13998 (DHFR-TS) [44]	K13938 (folM) []	K01938 (fhs) [9]	K13402 (MTHFD1L) [17]	K00288 (MTHFD) [113]	K01491 (folD) []	K13403 (MTHFD2) [57]
hsa P	1719				25902	4522		10797 441024
ptr P	460535 735442				463073	452965		461253
pon P	100437501 100455969				100437887	100173036		100442117
mcc P	711268				705222	707059		702907
mmu P	13361				270685	108156		17768 665563
rno P	24312				361472	64300		680308 313410 305248
cfa P	609048				476245	480352		483107 482197
aml P	100472337				100477350	100467466		100463985 100475974
bta P	508809				534296	534382		517539 536269
ssc P					100154722	414382		100525706
ecb P	100073256				100063441	100062154		100053990 100056425
mdo P	100012354				100032106	100027684		100024210 100023851
oaa P								
gga P	427317				421633	423508		426126 770327
mgp P	100540725				100542162	100542376		100538988
tgu P	100229812				100224764	100218066		100224318
acs P	100559043					100553949		100566586 100553649

图3-12 人类One carbon pool by folate通路的Ortholog table

在这个表中,行与物种对应,3个字母都是相应物中的英文单词缩写,比如has表示Homo sapiens, mcc表示Macaca mulatta; 列就表示相应的Ortholog分类。如上图同一物种后有多个条目,则表示在该物种中存在多个蛋白,它们分别由以上数字代表的基因所编码,空白则表示在该物种中不存在这种酶。

点击K00287则这一KO分类信息及成员列表都可显示出来; 点击has则链接到物种(人类)基因组去了; 点击P,则显示相应的代谢通路。下面我们点击1719,如图3-13所示:

KEGG Homo sapiens (human): 1719 Help

Entry	1719 CDS H.sapiens
Gene name	DHFR, DHFRP1, DYR
Definition	dihydrofolate reductase (EC:1.5.1.3)
Orthology	K00287 dihydrofolate reductase [EC:1.5.1.3]
Pathway	hsa00670 One carbon pool by folate hsa00790 Folate biosynthesis hsa01100 Metabolic pathways
Drug target	Methotrexate: D00142 D02115 Trimetrexate: D06238 D06239
Class	Metabolism; Metabolism of Cofactors and Vitamins; Folate biosynthesis [PATH:hsa00790] Metabolism; Metabolism of Cofactors and Vitamins; One carbon pool by folate [PATH:hsa00670] BRITE hierarchy
SSDB	Ortholog Paralog GFIT
Motif	Pfam: DHFR_1 PROSITE: DHFR Motif
Other DBs	NCBI-GI: 4503323 NCBI-GeneID: 1719 OMIM: 126060 HGNC: 2861 HPRD: 00519 Ensembl: ENSG00000228716 UniProt: P00374 B0YJ76

All links

Ontology (4)
KEGG BRITE (4)
Pathway (3)
KEGG PATHWAY (3)
Disease (2)
OMIM (2)
Drug (4)
KEGG DRUG (4)
Chemical reaction (7)
KEGG ENZYME (1)
KEGG REACTION (6)
Genome (1)
KEGG GENOME (1)
Gene (12)
KEGG ORTHOLOGY (1)
NCBI-Gene (1)
NCBI-GI (5)
UniGene (2)
HGNC (1)
HPRD (1)
ENSEMBL-HSA (1)
Protein sequence (6)
UniProt (2)
RefSeq(pep) (1)
IPI (3)
DNA sequence (11)
RefSeq(nuc) (1)
GenBank (5)
EMBL (5)
3D Structure (52)
PDB (52)
Protein domain (2)
Pfam (1)
PROSITE (1)
All databases (104)

图3-13 KEGG中的人类1719基因

如上图,就是我们常见的一个页面,1719是KEGG中的基因ID, H.sapiens表示物种,然后是基因的名称,表达的酶,属于哪个KO分类以及参与哪些代谢途径; 下面还有结构、序列信息等。所以从Ortholog table中可以很容易知道一张代谢通路上有哪些KO分类(酶类),并且这些酶类的成员在各物种中分配存在的情况以及特定的名称。

(三) KEGG数据库本地化及应用接口

KEGG提供了ftp服务，可以下载所有的数据，方便数据库的本地化。此外，KEGG还提供应用程序接口（application programming interface，API），利用KEGG API，用户可以方便的建立自己的客户端，从而获得最新的数据。KEGG API支持多种编程语言，如Perl、Java等，对操作系统和对象模块的选择也没有倾向性，表3-2列出了常用的KEGG API函数。

表3-2 常用的KEGG API函数

函数	功能
list_pathways	返回指定物种的所有通路
get_genes_by_enzyme	返回指定物种中编码指定酶的所有基因
get_ enzymes _by_gene	返回指定物种中指定基因编码的所有酶
get_enzymes_by_compound	返回所有物种中催化指定化合物的酶
get_enzymes_by_reaction	返回所有物种中催化指定反应的酶
get_ko_by_gene	返回指定基因对应的所有ko代码
get_genes_by_pathway	返回指定通路中的所有基因
get_compounds_by_pathway	返回指定通路中的所有化合物
get_reactions_by_pathway	返回指定通路中的所有反应

三、其他常见生物学通路数据库

Biocarta 数据库（http: //www.biocarta.com/）是较常用的通路数据，可以用来研究分子互作关系、富集分析、通路为基础的研究等。从分子的关系角度描绘了一个网络图模型。是“开源”数据库的典型代表法，通过社区论坛而成长起来的数据库，通过不断整合蛋白质组信息迅速发展壮大起来。它还提供了目录并且总结了12万多个多物种的基因信息的重要资源。发现了过去的已有的通路的同时也发现了一些新的通路。其中，Biocarta 是目前覆盖范围最广的信号通路数据库，包含了大量的通路细节知识，方便进行单个分子的查询，但是单个通路规模较小，不提供批量下载。人类生物学反应及信号通路数据库Reactome（http: //www.reactome.org）是一个汇集了由专家撰写，经同行评阅的有关人体内各项反应及生物学路径的文章的数据库，该数据库相当于一个有效的数据资源以及电子图书。该数据库为人们提供了一个全新的从整体水平上对生物学途径进行研究的工具，同时，它也是一个改良的搜索及数据挖掘工具，可以简化与生物学途径相关的数据搜索与研究。此外，对用户提供的高通量数据组进行分析，也变得更为简单。目前，由于直系同源预测方法的改进，反应组学数据库也开始收录其他模式生物的数据了，现在通过与其他数据库合作和人工注释方式，已经收录了包括拟南芥（Arabidopsis）、水稻（Oryza sativa）、果蝇（Drosophila）及原鸡（Gallus gallus）等22种模式物种的反应组学数据。反应组学的数据库内容和相关软件都是开源共享，免费使用的。Reactome作为经典的通路数据库建立时间较早，图示清楚，下载方便，但与Biocarta相比包含的通路数据不够全面，STKE数据库由通路专家进行收集整理，包括通用的细胞信号数据和部分组织细胞中特殊的信号过程，具有内容较详细但通路数目较少的特点。AfCS

数据库以信号分子为基础,提供其参与的相互作用及信号通路图,包含了 AfCS 项目最新的研究成果。而 Pathway Interaction Database 专门收集人的信号通路, 包含了大量文献挖掘得到的信号通路,并且从 Biocarta 和Reactome 中导入了大部分的信号通路,适于人的信号通路分析。此外, AMAZE 数据库采用专门的数据模型,可将单个生物分子和相互作用整合进细胞过程。其他常用的信号和代谢通路数据库详见表3-3。

表3-3 其他常用的信号和代谢通路数据库

数据库	网址	描述
PID	http: //pid.nci.nih.gov	文献挖掘的人信号通路数据库
STKE	http: //stke.sciencemag.org	参与信号转导的分子及其相互作用关系的信息
AfCS	http: //www.signaling-gateway.org	参与信号通路的蛋白质相互作用和信号通路图
AMAZE	http: //www.amaze.ulb.ac.be	对细胞过程的相关信息进行、注释和分析
BIND	http: //www.bind.ca	提供参与通路的分子的序列和相互作用信息
DOQCS	http: //doqcs.cbs.res.in	细胞信号通路的量化数据库
SigPath	http: //sigpath.org	提供细胞信号通路的量化信息
MetaCyc	http: //biocyc.org/metacyc/	微生物为主的多个物种的酶和代谢途径数据库
EcoCyc	http: //biocyc.org/ecocyc/	大肠埃希菌(K-12)基因组、基因产物和代谢通路
UM-BBD	http: //www.labmed.umn.edu/umbbd/	微生物生物催化反应和生物降解通路

第三节

基因功能预测

Section 3 Gene Function Prediction

一、基因功能预测的目的和意义

目前，大量参与重要生命活动的基因功能仍然未知。因此，生物信息学的重要任务之一是在全基因组范围内对基因功能进行预测。传统的基因功能预测方法主要依赖于序列的同源性，而近来已经发展了很多基于GO数据库或KEGG数据库的方法，利用高通量的基因表达和蛋白质互作数据进行功能预测，其中一些新开发的方法试图整合多种数据类型，通过构建功能相关网络的方式预测基因功能（图3-14）。GO数据库包含了基因参与的生物过程，所

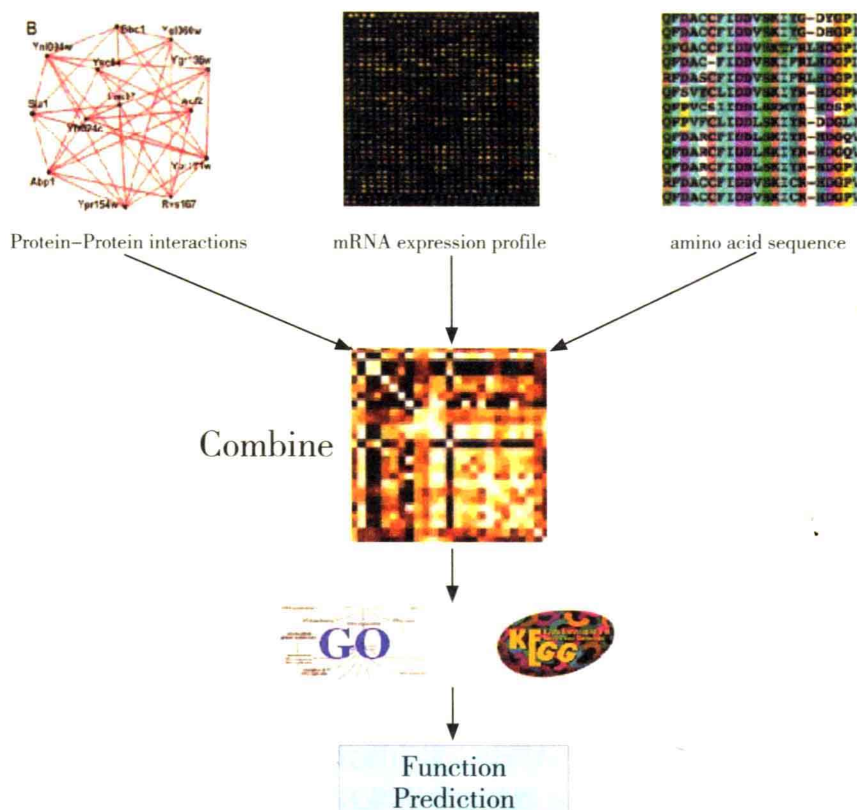


图3-14　整合蛋白质互作数据、表达谱和序列数据的功能预测

处的细胞位置及具有的分子功能三方面功能信息,通过GO中的注释信息,可以对基因的功能进行预测。KEGG是系统分析基因功能、联系基因组信息和功能信息的知识库, KEGG的PATHWAY 数据库提供了基因编码的生物学大分子酶或者蛋白质在生命体内相互联系相互影响的情况。同一生物学通路内的基因大多参与了此代谢通路所揭示的生命过程。根据功能相似的基因可能导致相似的表型这一依据,可以通过网络拓扑性质对基因的功能进行预测,并利用GO和KEGG功能富集分析方法进行进一步的预测。当前基于GO或KEGG的基因功能预测策略一般为: 首先,从总体上宏观地概括抽取信息,如不同样本间、不同时间点间全部差异基因; 其次,通过GO或KEGG分析,即从GO分类结果找到实验涉及的显著功能类别或将差异基因映射到通路中,根据基因在通路中的位置及表达水平的变化算出受影响显著的通路,从而预测未知的基因功能。

二、基因功能预测的基本原理

基于GO或KEGG的基因功能预测通常需要定义基因集,基因集的定义基于统一的先验生物学知识,如已发表的有关基因共表达、生物通路等。一个基因集是基因芯片上一组具有相同生物学功能或位于同一生物通道的基因,产生基因集的数据包括基因表达谱数据和蛋白质互作数据。

(一) 基于GO的基因功能预测

1. 对差异表达基因进行功能预测

GO应用的一个重要方面就是用来指导基于基因表达谱数据的基因功能预测。在基因芯片的数据分析中,研究者可以找出哪些差异表达基因属于一个共同的GO功能分支,并用统计学方法检验结果是否具有统计学意义,从而得出差异表达基因主要参与了哪些生物功能。

目前,大量的基因功能预测方法利用GO作为功能分类的来源或结果证实。在已知的大多数相关研究中,研究者首先将感兴趣基因注释到GO上,然后筛选出显著性富集的GO结点作为功能标签,考察这组基因是否共同注释到同一个功能结点上,或注释的结点是同一个结点的直接子结点,并认为这样的基因具有相似的功能,这项工作实现了对未知基因功能预测,是GO结构信息的进一步发掘。这是直接利用GO注释的方法进行基因功能预测。

目前许多已知功能的基因只注释到了描述很不具体的功能类,称之为已知部分功能的蛋白质。显然寻找这些基因的精细功能对于了解这些基因和提供必要的数据来学习其他基因的功能都具有重要意义。为了寻找已知部分功能的基因更精细的功能,目前有一种深层预测算法: 该算法利用蛋白质互作数据,将基因从其已注释到的功能类向下预测一层或多层,发现其更精细的功能。由于已知部分功能的基因参与一个子功能类的先验概率增大,预测的可靠性可能会提高,因此使用注释到同一个功能类中的基因可以过滤部分假阳性互作。

具体做法为: 首先,选定一个GO结点作为深层预测的目标结点,定义它的任何一个祖先结点为预测空间,按照GO注释的提示,将注释到预测空间而没有注释到它的任何一个子结点的基因定义为已知部分功能的基因,即预测对象; 然后,通过连接注释在预测空间中互作的基因构建一个功能特异的子网,孤立的基因被排除在外,在互作子网中,注释到目标结点

的蛋白质被当作阳性样本，而除预测对象外的其他蛋白质被当作阴性样本。

通常一个蛋白质被赋予与其直接相互作用的邻居蛋白质中出现频率最高的几个功能。尽管一个蛋白质可以执行多个功能，这里选择只为蛋白质赋予一个可信度最高的子功能。因为目标结点中阳性样本要和预测空间中所有其他子结点的阴性样本竞争，因此修改大数法对于预测一个阳性结果来说是保守的。可以采用留一法来评价分类器的预测效果。每一个训练样本都要被轮流留出来作为测试样本。计算真阳性(TP)、真阴性(TN)、假阳性(FP)和假阴性(FN)，再计算精确度、覆盖率和F指标。基于蛋白质互作数据和深层预测方法，以高于90%的精确率，为几千个已知部分功能的酵母和人类蛋白质预测了精细的功能。预测的精细功能对于指导随后的实验和提供必要的功能知识来学习其他蛋白质的功能都具有重要的意义。

2. 蛋白质互作网络用于基因功能预测

传统的基因功能注释及预测方法是根据基因相关的一些统计特征集，利用机器学习方法来得出功能注释的规则用于预测。基因功能实现的复杂性以及功能定义的模糊性，使得传统的利用特征预测的方法很难准确地进行预测。而蛋白质相互作用网络能够利用蛋白质之间的相关性，对未知功能的基因进行注释。目前，利用相互作用网络进行功能注释主要有两种方法，即直接注释方法(direct annotation schemes)和基于模块的方法(module assisted schemes)。

(1)直接注释方法：直接注释方法根据网络中某个蛋白质的连接情况直接推测该蛋白质的功能。这类方法基于的假设是：在蛋白质相互作用网络中，距离相近的两个蛋白质更加倾向于拥有相似的功能。而通过两蛋白质在网络中的距离来计算并判断这两个蛋白质功能相似性有许多的方法：①邻居结点计算法(neighborhood counting)：这种方法是最简便也是相对较早出现的方法。它根据网络中某个蛋白质直接相关的邻居已知蛋白质的功能来确定该未知蛋白质的功能注释。这种方法假设某未知蛋白质的邻居中有超过n个蛋白质具有一样的功能，就将这种功能赋予该蛋白质。这种方法虽然简单并且有时候非常有效，然而它在功能注释过程中不能为这种关联性提供非常有显著意义的解释，并且它也没有考虑到网络的全局拓扑结构。②图论方法(graph theoretic method)：图论方法不同于邻居结点计算法，它可以考虑网络的全局拓扑结构，基本思路是：对一个未知功能蛋白质赋予某种功能，要使得注释为相同功能的蛋白质(未注释或者已注释)的连接数目最多。③马尔可夫随机场方法(Markov random field method)：注释方法中有许多基于概率的方法，它们均基于马可夫假设：蛋白质的功能独立于与其直接相邻的邻居之外的所有蛋白质。根据这个假设，人们也提出了马可夫随机场模型用于蛋白质功能的注释。

(2)基于模块的方法：基于模块的方法首先将网络相关的蛋白质组成不同的模块，然后根据该模块中成员的功能来得到整个模块所共有的可能的功能，从而用来预测其中未知成员的功能。一个功能模块指其中的蛋白质所处的细胞位置以及相互作用使得它们可以实现一个特定的功能。而基于功能模块的蛋白质功能注释方法也不再单独预测单个蛋白质的功能，而是试图发现模块中所有蛋白质的共同内在的功能。一旦模块确定，那么可以通过一些简单的方法来预测其功能，比如该模块中如果大部分的蛋白质都具有某种功能，那么这种功能就将赋予该模块。对蛋白质相互作用网络进行模块划分的常用方法有以下几种：①分级聚类方法(hierarchical clustering based methods)：聚类就是将相似功能的蛋白质归为同一类

（模块）。分级聚类的关键问题是如何评判蛋白质对之间的相似性，最简单的方法是以两个蛋白质之间的距离作为基准。但是在分级聚类中，大量蛋白质对之间的距离都是相同的，通常认为同一个模块中的蛋白质成员更加可能拥有最短的路径距离谱（path distance profiles）。根据这个假设，所有短路径的蛋白质对聚成一类。这个方法实施比较复杂，很难在整个基因组水平上的网络上进行分析，但在一些子网络中它已经得到很好的应用，比如对酿酒酵母的核蛋白的相互作用网络分析。②图形聚类方法（graph clustering methods）：大量的图形聚类方法也用于图形化描述二元相互作用。早期的图形聚类方法用于相互作用网络模块的构建主要有两类，一类是基于SPC聚类（super paramagnetic clustering）方法，另一类为基于蒙特卡洛算法（monte carlo algorithm）。其中SPC算法在决定那些内部密度很高但松散的连接于其他部分的模块效果非常好。在最近，又不断发展出许多新的图形聚类算法，如高连通子图算法（highly connected sub graphs，HCS）、有限邻居搜索聚类算法（restricted neighborhood search clustering，RNSC）以及马尔可夫聚类算法（markov clustering，MCL）等。

（二）基于KEGG通路分析的基因功能预测

通路分析是现在经常被使用的芯片数据基因功能分析法。与GO分类法（应用单个基因的GO分类信息）不同，通路分析法利用的资源是许多已经研究清楚的基因之间的相互作用，即生物学通路。研究者可以把表达发生变化的基因集导入通路分析软件中，进而得到变化的基因都存在于哪些已知通路中，并通过统计学方法计算哪些通路与基因表达的变化最为相关。现在已经有丰富的数据库资源帮助研究人员了解及检索生物学通路，对芯片的结果进行分析。主要的生物学通路数据库有以下两个：①KEGG数据库：迄今为止，KEGG数据库是向公众开放的最为著名的生物学通路方面的资源网站。在这个网站中，每一种生物学通路都有专门的图示说明。②BioCarta数据库：BioCarta是一家生物技术公司，它在其公共网站上提供了用于绘制生物学通路的模板。研究者可以把符合标准的生物学通路提供给BioCarta数据库。BioCarta数据库不会检验这些生物学通路的质量，因此其中的资源质量参差不齐，并且有许多相互重复。然而BioCarta数据库数据量巨大，且不同于KEGG数据库，包含了大量代谢通路之外的生物学通路，所以也得到广泛的应用。

芯片数据通路分析的第一步是差异基因的通路定位（图3-15），一些商业软件如Genespring可以做到，基于EASE算法的开放在线程序DAVID也可以实现定位。目前的通路分析方法还存在很多局限性，例如只注意到基因集合定位到了哪个通路而忽略了其在通路中的位置，如果一个通路由某个基因产物触发或被单个受体激活，并且特定的蛋白没有表达，这个通路就会受到严重影响甚至关闭；相反，如果多个基因与某个通路相关但都只出现在通路的下游，那么其表达水平的变化就可能不会对通路造成很大影响。另外，一些基因往往有多个功能分布于不同的通路发挥不同的作用，要得到相对准确的结果还必须考虑通路的拓扑结构。目前很少有能将基因差异表达值变化应用于通路分析的方法，Pathwayexpress提出了一种基于IF（impact factor）的通路分析方法，综合了差异基因的标化的差异表达值、通路中基因的统计学显著性以及信号通路的拓扑学三方面内容。Pathwayexpress主要基于KEGG库，结果输出中自动把差异基因以不同颜色定位于通路中，红色为上调，蓝色为下调，这些定位着上调和下调基因的通路图可以在Java控制台中找到绝对路径，在浏览器中打开或保存，也可以GML格式导出，然后直接导入Cytoscape，用merge结点功能把多个相关pathway连接起来，显

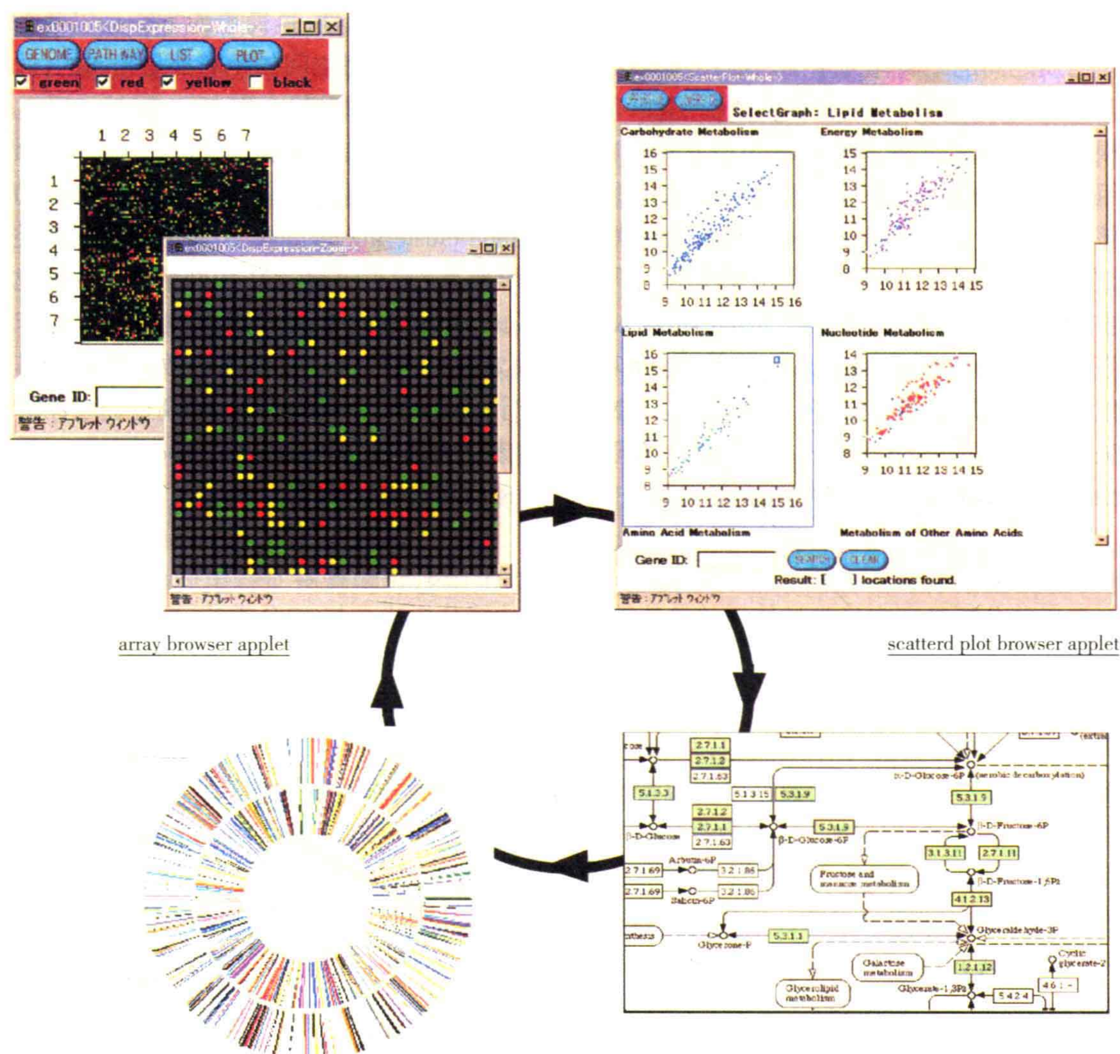


图3-15 通过表达谱数据进行通路定位

示互作网络，并分别以红蓝色显示显著性通路中上调下调的基因（结点），以及这些基因与其他基因间的相互作用（边），可以从不同视角观察其位置，不断放大就可以看到结点的基因名称。其他的可视化工具还有pathwaystudio、genmapp、arrayxpath、osprey等。Biolayout也是一款分子作用网络展示工具，所不同的是结果为三维图形界面。

三、基因功能预测的常用工具

（一）基于GO的基因功能分析软件

EASE（expressing analysis systematic explorer）是比较早的用于芯片功能分析的网络平台。由美国国立卫生研究院（NIH）的研究人员开发。研究者可以用多种不同的格式将芯片中得到的基因导入EASE 进行分析，EASE会找出这一系列的基因都存在于哪些GO分类中。其

最主要特点是提供了一些统计学选项以判断得到的GO分类是否符合统计学标准。EASE能进行的统计学检验主要包括Fisher 精确概率检验，或是对Fisher精确概率检验进行了修饰的EASE 得分（EASE score）。

由于进行统计学检验的GO分类的数量很多，所以EASE采取了一系列方法对“多重检验”的结果进行校正。这些方法包括Bonferroni校正法、Benjamini falsediscovery rate和bootstraping。同年出现的基于GO分类的芯片基因功能分析平台还有底特律韦恩大学开发的Onto-Express。2002年，Norway大学和Uppsala大学联合推出的Rosetta 系统将GO分类与基因表达数据相联系，引入了“最小决定法则”（minimal decision rules）的概念。它的基本思想是在对多张芯片结果进行聚类分析之后，与表达模式不相近的基因相比，相近的基因更有可能参与相同的生物学功能的实现。比较著名的基于GO分类法的芯片数据分析网络平台还有很多，这里列举了其中的一部分（表3-4）：

表3-4　用GO 分类法进行芯片功能分析的网络平台

平台名称	网址
Onto-Tools	http: //vortex.cs.wayne.edu/projects.htm
ROSETTA	http: //rosetta.lcb.uu.se/general/
GOToolBox	http: //burgundy.cmmt.ubc.ca/GOToolBox/
GOstat	http: //gostat.wehi.edu.au/
GFINDer	http: //www.medinfopoli.polimi.it/GFINDer/
FatiGO	http: //www.fatigo.org/
EASE	http: //david.abcc.ncifcrf.gov/ease/ease.jsp

（二）基于KEGG的基因功能分析软件

最先出现的通路分析软件之一是GenMAPP（gene microarray pathway profiler），它可以免费使用，其最新版本为Gen-MAPP2。在这个软件中，使用者可以用几种灵活的文件格式输入自己的表达谱数据，GenMAPP的基因数据库包含许多从常用的资源中得到的物种特异性的基因注释和识别符（ID）。这些ID可以将使用者输入的基因与不同的生物学通路的基因联系起来。这些生物学通路存在于GenMAPP 的MAPP文件中。MAPP文件需要时常下载更新。它包含有许多KEGG生物学通路，一些GenMAPP自己的生物学通路和许多GO分类的MAPP文件，全部操作简单明了。而且依靠其自带的MAPPBuilder和MAPPFinder 两个软件，使用者可以自己绘制生物学通路和对MAPP 文件进行检索。由于使用者可以自己绘制生物学通路保存为MAPP格式，而且这个文件很小，易在网络上传播，所以GenMAPP数据库更有利于研究者之间的及时交流。由于上述特点，GenMAPP数据库及软件仍是现今免费平台里应用比较广泛的。

2004年发表的Pathway Miner也是应用较为广泛的免费通路分析网络平台，由美国亚利桑那大学癌症中心建立维护，其最突出的特点就是信息全面，操作简便。使用者可以在这个网站中获得单个基因的序列、功能注释，以及有关它们编码的蛋白结构功能，组织分布，OMIM等信息。对于通路分析部分，使用者给出基因集及它们的表达变化值，网站可以根据

三大公用的通路数据库：KEGG、GenMAPP和BioCarta，生成变化基因参与的通路，并用Fisher精确概率检验。PathwayMiner自动把得到的通路分成两大类：代谢通路和细胞调节通路。方便使用者根据不同的研究目的选择需要查看的结果。在2006年国内也开发了用于通路分析的网络平台，即KOBAS（KO-Based annotation system），其基于KEGG数据库建立，由北京大学生命科学院开发和维护。其特点是可直接采用基因或蛋白质的序列录入基因，并对录入的基因集进行KO注释。对于结果的可靠性检验提供了四种统计方法。使用者可以在网站进行注册，网站会为使用者保存输入的数据，方便日后直接调用。最近推出的软件Eu.Gene整合了来自KEGG、Gen-MAPP以及Reactome的通路数据，并采用Fisher精确概率检验及基因集富集分析（Gene Set Enrichment Analysis，GSEA）来检验结果是否具有统计学意义。这里列举了部分通路分析的网络平台及它们的网址（表3-5）。

表3-5 通路分析网络平台

平台名称	网址
GenMAPP	http://www.genmapp.org/
PathwayMiner	http://www.biorag.org/pathway.html
KOBAS	http://kobas.cbi.pku.edu.cn
GEPAT	http://gepat.bioapps.biozentrum.uni-wuerzburg.de/GEPAT/index.faces
VitaPad	http://bioinformatics.med.yale.edu/group
KEGGanim	http://biit.cs.ut.ee/kegganim/
WholePathwayScope	http://www.abcc.ncifcrf.gov/wps/wps_index.php
VisANT 3.0	http://visant.bu.edu/
Eu.Gene	http://www.ducciocavalieri.org/bio/Eugene.htm

第四节

基因集合富集分析

Section 4 Gene Set Enrichment Analysis

一、富集分析的目的和意义

已建立的基因及其产物注释数据库包含了丰富的知识和复杂的结构，促使研究人员开展以注释数据库为知识基础的基因功能研究，以便更好地利用注释系统。一组基因直接注释的结果是得到大量的功能结点。这些功能具有概念上的交叠现象，导致分析结果冗余，不利于进一步的精细分析，所以研究人员希望对得到的功能结点加以过滤和筛选，以便获得更有意义的功能信息。目前最常用的方法是基于GO或KEGG的富集分析。人们通过多种方法获得大量的感兴趣基因，如差异表达基因集、共表达基因模块、蛋白复合物基因簇等，然后寻找这些感兴趣基因集显著富集的GO结点或KEGG通路，这有助于指导进一步深入细致的实验研究。

二、富集分析的基本原理

传统的单基因分析方法存在许多缺陷，如难以对芯片分析中筛选出大量的差异表达基因合理的解释、未考虑基因间相互作用、不能有效地利用一些先验信息、差异表达基因可重复性差等问题。为了克服单基因分析的诸多缺点，提出了基于已定义的基因集（gene set）进行分析的方法——基因富集分析（gene set enrichment analysis，GSEA）。基因集的定义基于统一的先验生物学知识，如已发表的有关生物通道、基因共表达信息等。一个基因集是指一组具有相同生物学功能或位于同一生物通道的基因。最常用于基因集的基因注释数据库有Gene Ontology（GO）和KEGG。一组基因直接注释的结果是得到大量的功能结点。这些功能具有概念上的交叠现象，导致分析结果冗余，不利于进一步的精细分析，所以研究人员希望对得到的功能结点加以过滤和筛选，以便获得更有意义的功能信息。目前最常用的方法是基于GO或KEGG的富集分析。人们通过多种方法获得大量的感兴趣基因，如差异表达基因集、共表达基因模块、蛋白复合物基因簇等，然后寻找这些感兴趣基因集显著富集的GO结点或KEGG通路，这有助于指导进一步深入细致的实验研究。

因此，富集分析方法通常是分析一组基因在某个功能结点上是否过出现（over-presentation）。这个原理可以由单个基因的注释分析发展到大基因集合的成组分析。由于分析的结论是基于一组相关的基因，而不是根据单个基因，所以富集分析方法增加了研究的可靠性，同时也能够识别出与生物现象最相关的生物过程。富集分析中常用的统计方法有累计超几何分布、Fisher精确检验等。

累计超几何分布公式：

$$P(X>q)=1-\sum_{x=1}^{q}\frac{\left(\frac{n}{x}\right)\left(\frac{N-n}{M-x}\right)}{\left(\frac{N}{M}\right)} \qquad (3\text{-}1)$$

其中N为注释系统中基因总数，n为将要考察的结点或通路本身注释的基因数，m为感兴趣的基因集大小，x为基因集与结点或通路的交集数目。

Fisher精确检验公式：

$$P=\frac{\left(\frac{a+b}{a}\right)\left(\frac{c+d}{c}\right)}{\left(\frac{n}{a+c}\right)} \qquad (3\text{-}2)$$

n为系统中基因总数，a为感兴趣的基因集中的基因数目，b为将要考察的结点或通路本身所注释的基因数目，c为去除感兴趣基因以外的基因数目，d为待考察结点基因去除与感兴趣基因重合的数目。

三、富集分析常用工具

（一）GO富集分析常用工具

利用富集分析方法，对基因注释数据库做生物信息学研究产生了很多富集分析工具。这些工具对促进基因功能分析以及研究高通量的生物学数据起到了重要的作用。表3-6列举一些常用富集分析工具。在芯片的数据分析中，研究者可以找出哪些变化基因属于一个共同的GO功能分支，并用统计学方法检定结果是否具有统计学意义，从而得出变化基因主要参与了哪些生物功能。EASE是比较早的用于芯片功能分析的网络平台。由美国国立卫生研究院（NIH）的研究人员开发。研究者可以用多种不同的格式将芯片中得到的基因导入EASE进行分析，EASE会找出这一系列的基因都存在于哪些GO分类中。其最主要特点是提供了一些统计学选项以判断得到的GO分类是否符合统计学标准。EASE能进行的统计学检验主要包括Fisher 精确概率检验，或是对Fisher精确概率检验进行了修饰的EASE 得分（EASE score）。

续表

表3-6　常用GO分析的网络平台及网址

数据库	网址
ROSETTA	http: //rosetta.lcb.uu.se/general/
GOToolBox	http: //burgundy.cmmt.ubc.ca/GOToolBox/
Onto-express	http: //vortex.cs.wayne.edu/projects.htm
EASE	http: //david.abcc.ncifcrf.gov/ease/ease.jsp
GoMiner	http: //discover.nci.nih.gov/gominer/index.jsp
GOStat	http: //gostat.wehi.edu.au/
GFINDer	http: //www.medinfopoli.polimi.it/GFINDer/
g: Profiler	http: //biit.cs.ut.ee/gprofiler/
GOEAST	http: //omicslab.genetics.ac.cn/GOEAST/
GSEA	http: //www.broadinstitute.org/gsea/
DAVID	http: //david.abcc.ncifcrf.gov/

由于进行统计学检验的GO分类的数量很多，所以EASE采取了一系列方法对“多重检验”的结果进行校正。这些方法包括弗朗尼校正法（Bonferroni），本杰明假阳性率法（Benjamini falsediscovery rate）和靴带法（bootstraping）。同年出现的基于GO分类的芯片基因功能分析平台还有Wayne state大学开发的Onto-Express。2002年，挪威大学和乌普萨拉大学联合推出的Rosetta 系统将GO分类与基因表达数据相联系，引入了“最小决定法则”（minimal decision rules）的概念。它的基本思想是在对多张芯片结果进行聚类分析之后，与表达模式不相近的基因相比，相近的基因更有可能参与相同的生物学功能的实现。

（二）KEGG富集分析常用软件

通路分析是现在经常被使用的芯片数据基因功能分析法。与GO分类法（应用单个基因的GO分类信息）不同，通路分析法利用的资源是许多已经研究清楚的基因之间的相互作用，即生物学通路。研究者可以把表达发生变化的基因列表导入通路分析软件中，进而得到变化的基因都存在于哪些已知通路中，并通过统计学方法计算哪些通路与基因表达的变化最为相关。现在已经有丰富的数据库资源帮助研究人员了解及检索生物学通路，对芯片的结果进行分析（表3-7）。

表3-7　常用通路分析的网络平台及网址

数据库	网址
DAVID	http: //david.abcc.ncifcrf.gov/
GenMAPP	http: //www.genmapp.org/

续表

数据库	网址
PathwayMiner	http: //www.biorag.org/pathway.html
KOBAS	http: //kobas.cbi.pku.edu.cn
VitaPad	http: //bioinformatics.med.yale.edu/group
KEGGanim	http: //biit.cs.ut.ee/kegganim/
WholePathwayScope	http: //www.abcc.ncifcrf.gov/wps/wps_index.php
VisANT 3.0	http: //visant.bu.edu/

最先出现的通路分析软件之一是GenMAPP(gene microarray pathway profiler)。它可以免费使用,其最新版本为Gen-MAPP2。在这个软件中,使用者可以用几种灵活的文件格式输入自己的表达谱数据, GenMAPP的基因数据库包含许多从常用的资源中得到的物种特异性的基因注释和识别符(ID)。这些ID可以将使用者输入的基因与不同的生物学通路的基因联系起来。这些生物学通路存在于GenMAPP 的MAPP文件中。MAPP文件需要时常下载更新。它包含有许多KEGG生物学通路,一些GenMAPP自己的生物学通路和许多GO分类的MAPP 文件,全部操作简单明了。2004年推出的Pathway Miner也是应用较为广泛的免费通路分析网络平台,由美国亚利桑那大学癌症中心建立维护,其最突出的特点就是信息全面,操作简便。使用者可以在这个网站中获得单个基因的序列、功能注释,以及有关它们编码的蛋白结构功能,组织分布, OMIM等信息。对于通路分析部分,使用者给出基因列表及它们的表达变化值,网站可以根据三大公用的通路数据库: KEGG、GenMAPP和BioCarta,生成变化基因参与的通路,并用fisher 精确概率检验。PathwayMiner自动把得到的通路分成两大类: 代谢通路和细胞调节通路。方便使用者根据不同的研究目的选择需要查看的结果。在2006年国内也开发了用于通路分析的网络平台,即KOBAS(KO-based annotation system),其基于KEGG数据库建立,由北京大学生命科学学院开发和维护。其特点是可直接采用基因或蛋白质的序列录入基因,并对录入的基因列表进行KO 注释。对于结果的可靠性检验提供了四种统计方法。使用者可以在网站进行注册,网站会为使用者保存输入的数据,方便日后直接调用。最近推出的软件Eu.Gene 整合了来自KEGG, Gen-MAPP 以及Reactome 的通路数据,并采用fisher 精确概率检验及基因集富集分析(GSEA)来检验结果是否具有统计学意义。

四、富集分析应用实例

目前有很多方便易用的软件可以对基因集做富集分析,如DAVID, GO-2D, GOEAST等,都提供多种参数选择和丰富的结果分析。用户提交感兴趣的基因集,软件反馈给用户这组基因集富集在哪些结点上,每个结点注释的基因数目,统计检验的P值,并提供GO系统的可视化。

上面介绍了多种富集分析工具，这里以目前应用较为广泛的DAVID为例对基因集进行具体分析（图3-16）。DAVID是一个综合工具，不但提供基因富集分析，还提供基因间ID的转换、基因功能的分类等工具。

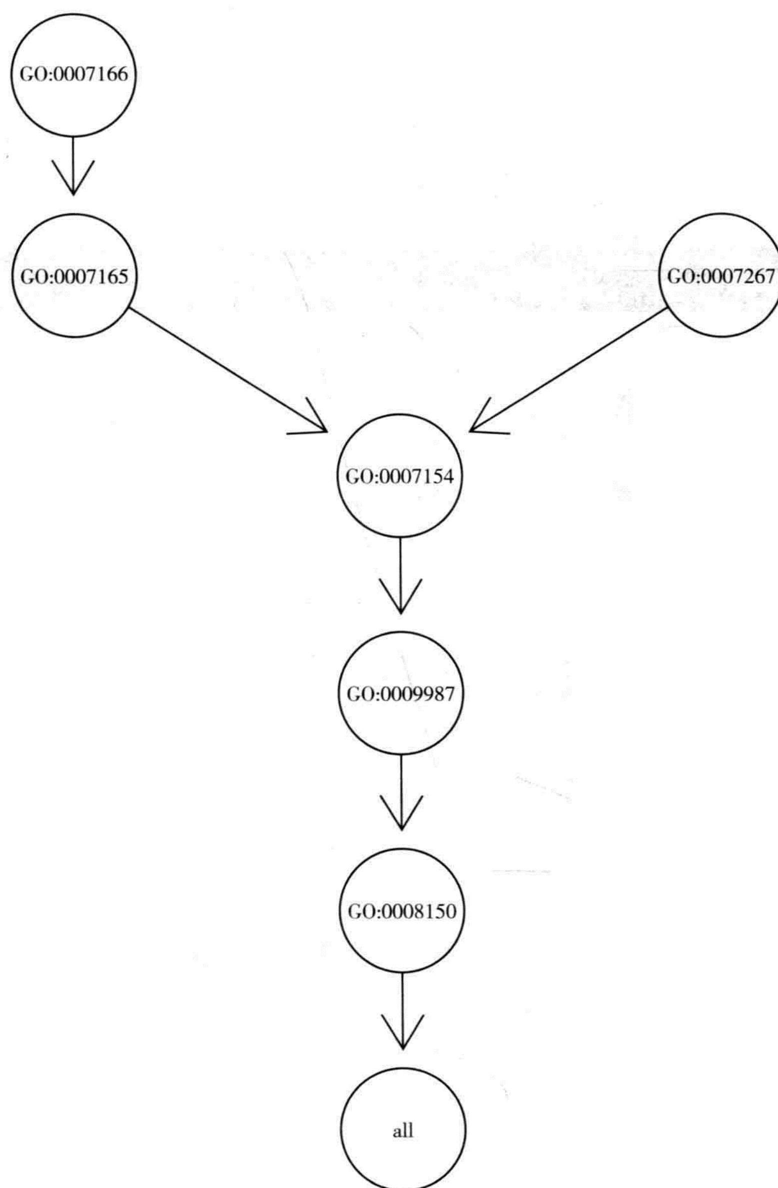


图3-16　DAVID工具应用首页

点击"Start Analysis"后，第一步为提交基因集，选择基因标识名和基因集类型；第二步得到注释结果摘要（图3-17），包括多种注释数据；然后选择感兴趣的注释内容得到富集分析结果。

这里以KEGG通路的富集分析为例（图3-18）。提交之后的结果如图3-18，可以看到，对提交的基因集做富集分析，找到5个具有显著性的通路。这里的"P-Value"是通过Fisher精确检验得到的p值，"Benjamini"指的是本杰明假阳性率校正方法。

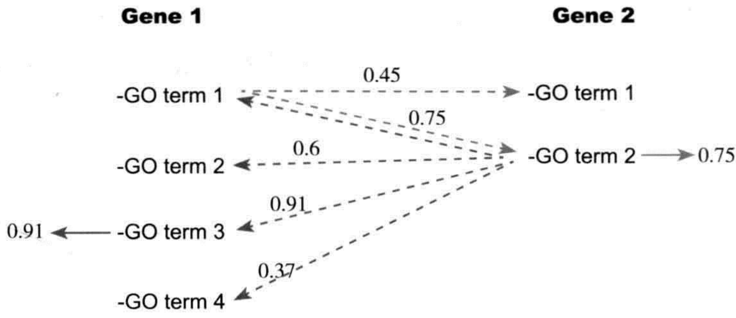


图3-17 DAVID富集分析注释结果摘要

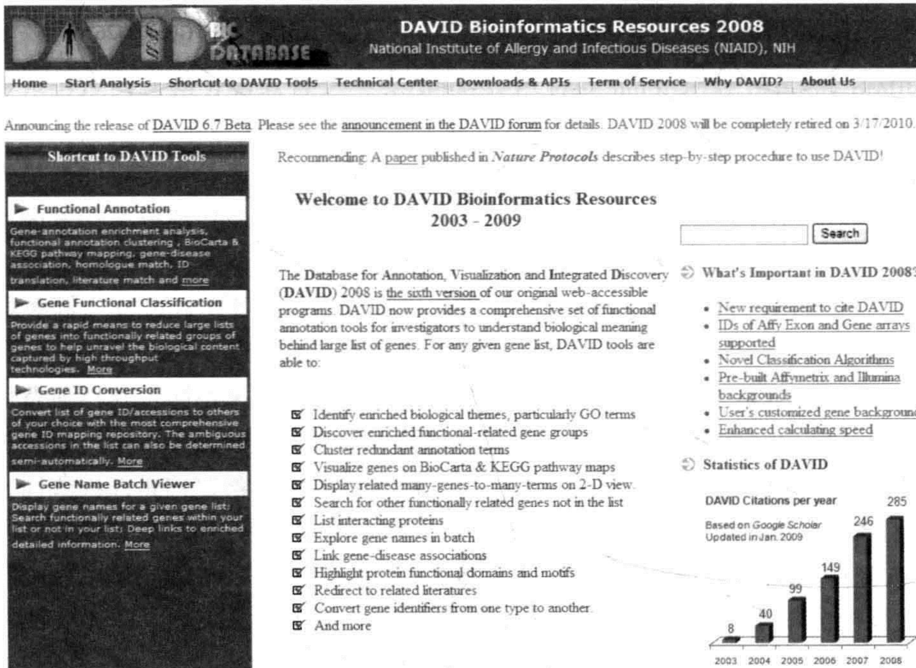


图3-18 DAVID在KEGG上富集结果实例

第五节 基因功能比较

Section 5 Gene Function Comparison

一、基因功能比较的意义

自从林奈提出了分类系统理论(system of taxonomy)和达尔文提出了生物进化论(theory of evolution),比较和分类的研究已经成为生物学的中心支柱。生物学不同于其他学科的原因是,其知识很少能够被减少到数学形式。因此,生物学家们希望能够利用自然语言或寻求其他的形式来组织知识体系来记录复杂的生物学知识。例如,在科学出版物中,在分类学计划书中,如何编辑、整理、展现生物体系知识。最基础的科学知识是生物规律和模式的比较,即实体间的比较,例如,比较基因、细胞、有机体、种群、物种等,从而发现它们的相似特征和差异特征。当出现新的实体时,生物学家可以通过比较他们来了解实体并根据他们间的相似程度进行多方面知识的推论。比较实体的方法已经越来越受到科学家们的重视。例如,两个基因的序列或结构可以直接(通过序列的对其比较算法)相比,同样的功能方面的比较也是如此,不同的是,序列和结构有一个客观的代表性和可衡量的特征,而功能方面却没有这样的特征,但这并不意味着功能比较必须在一个共同的和客观的形式表达上比较,所以功能比较并不是不可能的。

自动测序的出现对生物学知识的探索起到了深刻的影响。作为实验学的方法,研究的范围已经从基因水平转移到了基因组水平上,计算分析已经被证明在处理越来越多的数据时是必不可少的方法。因此,采取共同的和客观的知识表现方式,来帮助共享知识和计算机推理已成为关键。这种需求直接导致了本体的发展,如注释基因产物(基因本体论),注释序列(序列本体论),注释的实验分析的本体(基因芯片和基因表达数据的本体论)。

注释本体的应用提供了一种比较实体的手段。例如,如果两个基因产物具被注释在同一个体系中,那么我们可以比较它们所注释术语的相似性从而判断两个基因产物的相似性。虽然这种比较含蓄、间接(例如,找到一组基因产物相互作用的共同术语),但利用语义相似性的方法却可以得到一个明确的比较。语义相似性测度多年来一直是自然语言处理和信息检索研究的重要组成部分,是计算语言学和人工智能应用中亟待解决的问题。

在生物学中,基因本体论(gene ontology)主要集中在分子生物学中的语义相似性的研究,不仅因为它是生命科学界最广泛采用的本体,也因为它在比较基因产物的功能上的广泛应用。基于GO注释体系的语义相似性方法的应用为基因产物的功能比较提供了很好的出

口。GO应用的一个重要方面就是对G0术语的语义相似性进行度量。通过定义一个语义相似性测度,来度量两个本体或两个术语的相似性所返回的数值,反映了它们之间的亲密程度,这将大大提高基因研究工作的效率,节省更多的人力物力。

目前已经开发了很多基于GO等结构化数据库的基因功能相似性算法。这些算法对于构建功能网络以及预测基因功能有重要意义,成为生物医学研究与应用中的重要工具。对基因功能的比较可以了解未知基因的功能,认识基因与疾病的关系,掌握基因的产物及其在生命活动中的作用等。

二、语义相似性原理与算法

下面介绍如何利用信息理论体系中的相似性概念,来比较基因间的功能相似性。基于基因本体论,从特定蛋白质的功能信息出发,查找与其功能相似或者相关的蛋白质,或者对两个蛋白质之间的关联程度进行比较、量化,从而推测它们在生命活动中扮演的角色关系。通常认为,如果两个基因产物的功能相似,那么它们在GO中注解的功能术语就相近,所以我们只要能求出G0中术语对的相似度,就可以近似估计两基因产物功能的相似程度。通过研究,如果能找到新的计算语义相似度的方法,使GO术语间的语义相似度更加精确,那么就能更加精确地查找功能相似或者相关的蛋白质,从而更加精确地估计两基因产物功能的相似程度。

人们广泛了解的是Resnik在1995年提出的对分类系统中每个类定义的语义相似性算法,计算两个类的语义相似性,后有多位科学家经过改进等提供了多种类相似性的计算测度。在2002年Lord第一次提出把语义相似性理论应用到GO分类系统中,计算两个术语之间的相似性,从而可以利用不同的方法计算基因间的功能相似性,最后可以根据功能相似性得分预测未知基因的功能。

在GO这种层级结构的词汇分类系统中,从父术语到子术语,含义是逐层深入的关系。越往下层,概念越具体。换言之,越往下层,术语的信息含量越大,根术语的信息量近似为0。在分类系统中,利用GO结构信息和基因注释信息,首先设一个函数,计算得到每个术语的信息含量值: $p(c)=\dfrac{freq(c)}{N}$,freq(c)表示术语及它的子术语上注释的所有基因数,p(c)是术语c的概率,并且随着术语c在层级结构中的升级,概率p是单调递增的,top术语概率是1。则术语c的信息含量值为: $IC=-\log(p(c))$。由公式可知,术语的概率越大,而它的信息含量越小。即如果c1是c2的下属,则p(c1)≤p(c2)。所以说根术语的信息含量最小,越往下层,信息含量越大,即信息含量随着层级结构的深度增加更增大。这样本体体系中的每一个术语都被量化,都具有一个信息含量值,代表了这个术语所含有的信息量。所有方法术语间的比较和基因间的比较都是依靠这个信息含量值来进行进一步计算的。

三、生物学术语相似性

得到每个术语的信息含量值后,计算任意两个术语的相似性方法有多种,Resnik提出的语义相似性概念是定义为两个术语的公共祖先中最近距离的祖先术语的IC值即为它们的相

似性值，即

$$sim(c_1,c_2)=\max_{c\in S(c1,c2)}\left[-\log p(c)\right] \tag{3-3}$$

为了说明GO结构中结点关系，如何计算两个结点间共同祖先的最近祖先，如图3-19中解释GO: 0007154即为GO: 0007166和GO: 0007267的最近共同祖先。可以看到GO: 0007166，GO: 0007267的共同最近祖先即为GO: 0007154，也就认为它的IC值为两个结点的相似性值。在Resnik的方法中，若不同结点对的祖先相同，那么任何子层的结点对的相似性就没有区别，不能加以比较了，显然这是不合理的。Lin的方法与Resnik的信息量的方法有些相似，在理论上是很有根据的。这种方法的改进之处在于: 其一，两个要比较概念的信息量之和的标准化; 其二，假定要比较的两个概念是独立的。该方法把两基因产物的相似性定义为两术语共同的最近祖先术语的信息量与两术语平均信息量的比。

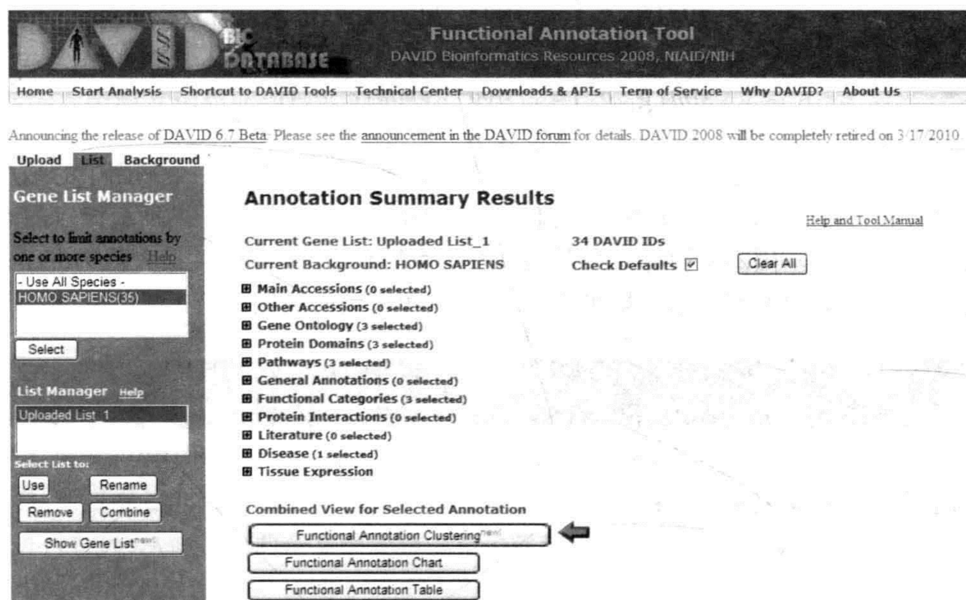


图3-19 GO结构示意图

GO: 0007154即为GO: 0007166和GO: 0007267的最近共同祖先

$$sim(c_1,c_2)=\frac{2IC_{ms}(c_1,c_2)}{IC(c_1)+IC(c_2)} \tag{3-4}$$

Jiang和Conrath的方法继承了图表中基于边的方法的特点，并且结合了基于术语的信息量的方法来计算术语对之间的相似度。但是也考虑了连接概念之间边的数目，还有局部密度，以及概念之间的连接类型等相关因素。这种方法尤其注意了连接父术语与子术语之间的边的连接强度。在上一种方法中我们已经讨论了子术语的实例概率与其父术语之间的关系，所以根据信息论我们整理得到:

$$sim(c_1,c_2)=2IC_{ms}(c_1,c_2)-\left[IC(c_1)+IC(c_2)\right] \tag{3-5}$$

以上几种方法在生物学研究中是比较常用的，也有一些其他的方法不断地被提出来。

比如基于语义路径覆盖的Combine算法，该算法首先计算出每个术语的信息量，然后分别计算两个术语的语义路径的交集之间术语信息量之和以及这两个术语语义路径的并集之间术语信息量之和，将这两者之间的比率作为相似性度量值，文中不做详细介绍。

四、基因（基因产物）功能比较

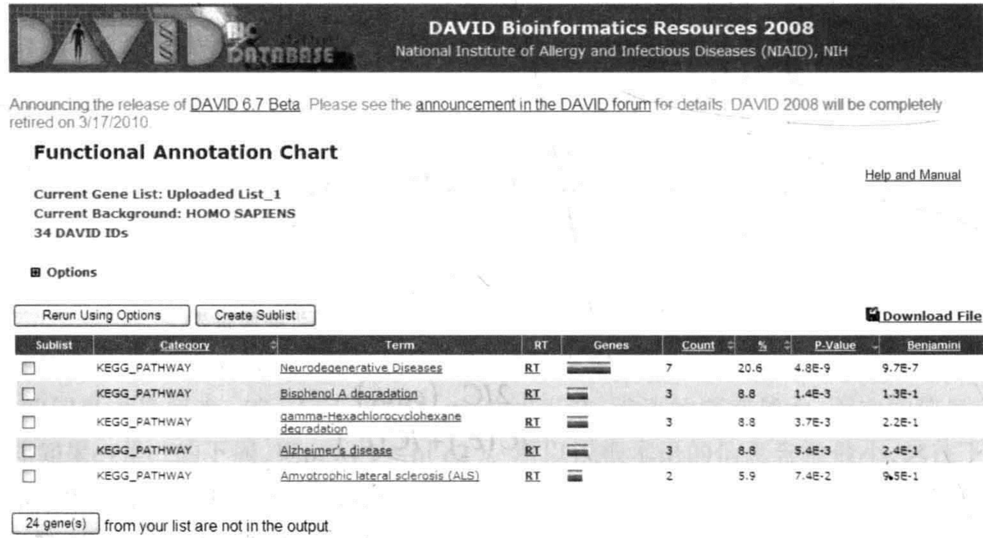
在GO系统中，可以计算得到任意两个术语的相似性值，则可根据基因注释在哪些术语上而计算两个基因之间的功能相似性。最简单的方法是取两个基因所注释的术语对的最大值或平均值，来作为两个基因的功能相似性。

对于给定的两个基因，它们的GO注释对应于术语集合 $c_i = c_1, c_2 \dots, c_N$ 和 $c_j = c_1, c_2 \dots, c_M$，则公式表示为：

$$sim(g_1, g_2) = \max_{1 \leqslant i \leqslant M, 1 \leqslant j \leqslant N} \left(sim\left(c_i, c_j\right) \right) \tag{3-6}$$

$$sim(g_1, g_2) = \underset{1 \leqslant i \leqslant M, 1 \leqslant j \leqslant N}{avg} \left(sim\left(c_i, c_j\right) \right) \tag{3-7}$$

最优分配法是目前被广泛应用的方法，如图3-20所示，首先取出一个基因中的结点与另一基因中的所有结点的语义相似性最大值，即基因1中结点1与基因2中的所有结点的语义相似性最大值为0.75，基因2中的结点2与基因1中所有结点的语义相似性最大值为0.91；分别计算出每个结点最大值，最后求和取平均值，即为两个基因的最优功能相似性值。公式如下：



Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Neurodegenerative Diseases	RT		7	20.6	4.8E-9	9.7E-7
	KEGG_PATHWAY	Bisphenol A degradation	RT		3	8.8	1.4E-3	1.3E-1
	KEGG_PATHWAY	gamma-Hexachlorocyclohexane degradation	RT		3	8.8	3.7E-3	2.2E-1
	KEGG_PATHWAY	Alzheimer's disease	RT		3	8.8	5.4E-3	2.4E-1
	KEGG_PATHWAY	Amyotrophic lateral sclerosis (ALS)	RT		2	5.9	7.4E-2	9.5E-1

图3-20 最优方法计算两个基因功能相似性示意图

$$rowScore = \frac{1}{N} \sum_{1}^{N} \max_{1 \leqslant j \leqslant M} S_{ij} \tag{3-8}$$

$$columnScore = \frac{1}{M} \sum_{1}^{M} \max_{1 \leqslant j \leqslant N} S_{ij} \tag{3-9}$$

$$sim(s1,s2)=\frac{rowScore+columnScore}{2} \tag{3-10}$$

五、基因集合功能比较

高通量实验在生物学领域的应用，得到了大量的基因集合数据，对这些基因集合数据的研究分析越来越被科学家们关注。这些基因集合常常被用来作为分子标记物去识别复杂疾病的遗传机制。它们通常是在特定的生物学条件下得到的，而对于那些具有相关但又不完全相同的条件下得到的不同的基因集，研究人员希望寻找它们之间的关联，例如对于同样本在不同实验平台下检测到的差异表达基因的可重复性等。仅仅利用基因集间的重复度作为测度来衡量它们之间的相似性已不能满足科学家的要求，科学家们不断发展和改良生物信息学方法去研究这一问题。

语义相似性的比较方法为这一方面的研究提供了可能。对基因集合找到其功能注释结点，从而利用语义相似性的理论对基因集合间进行功能比较，量化得分，从而实现了从生物学功能水平去比较基因集的功能相似性。无论是利用基因注释的方法还是基因集方法都可以找到这个基因集的功能术语集合，这已经在前面的章节进行了详细介绍。然后对于两个基因集合得到的功能术语基因进行语义相似度量，则是与基因间的相似性的计算方法相同，读者可以根据自己数据的特点选择不同的测度。

目前对于基因集合间的功能比较和量化，基于语义相似性方法研究者们已经开发了很多的方法可供利用，这里我们介绍两个常用的方法，一个利用单个基因间的功能相似性的整合分析，另一个是利用基因集的全局功能的整体分析（图3-21）。

第一个方法的思想是，对于两个基因集中寻找重复基因计数，然而大量的非重复基因无法计算，所以寻找非重复基因是否在功能上相关。利用GO和蛋白质互作网上的关联性对基因对进行打分，从而找出两个基因集合中相关联基因对的比例，作为判断两个基因集合功能是否相关的标准。

第二个方法是利用基因集的整体功能进行比较两个基因集的功能相似程度。这个方法基于GO对于每个基因集进行富集分析得到显著性术语，表示这个基因集的全局功能，再对这些术语按与基因集的相关程度加权。对两个基因集合得到的两个带有权重的术语集合做语义相似性计算，利用最佳匹配原则，可以算出它们的相似性得分。最后按相同数目的基因进行随机扰动，统计基因集的相似性得分是否显著，从而比较两个基因集是否功能相似。

六、常用工具

目前已经有一些比较基因间关联程度的算法和工具，利用语义相似性原理计算基因间功能相似性的工具已经有很多。我们以GOSim举例说明，GOSim是一个R包的工具。GOSim不但可以提供两个结点的语义详细性和两个基因间的功能相似性，还进行了进一步的功能分析，即基于基因在GO上的功能相似性对基因进行聚类，并对聚类结果提供可视化，这为研究者提供了大大的方便。比较著名的基于语义相似性的方法来做基因比较分析的工具还有

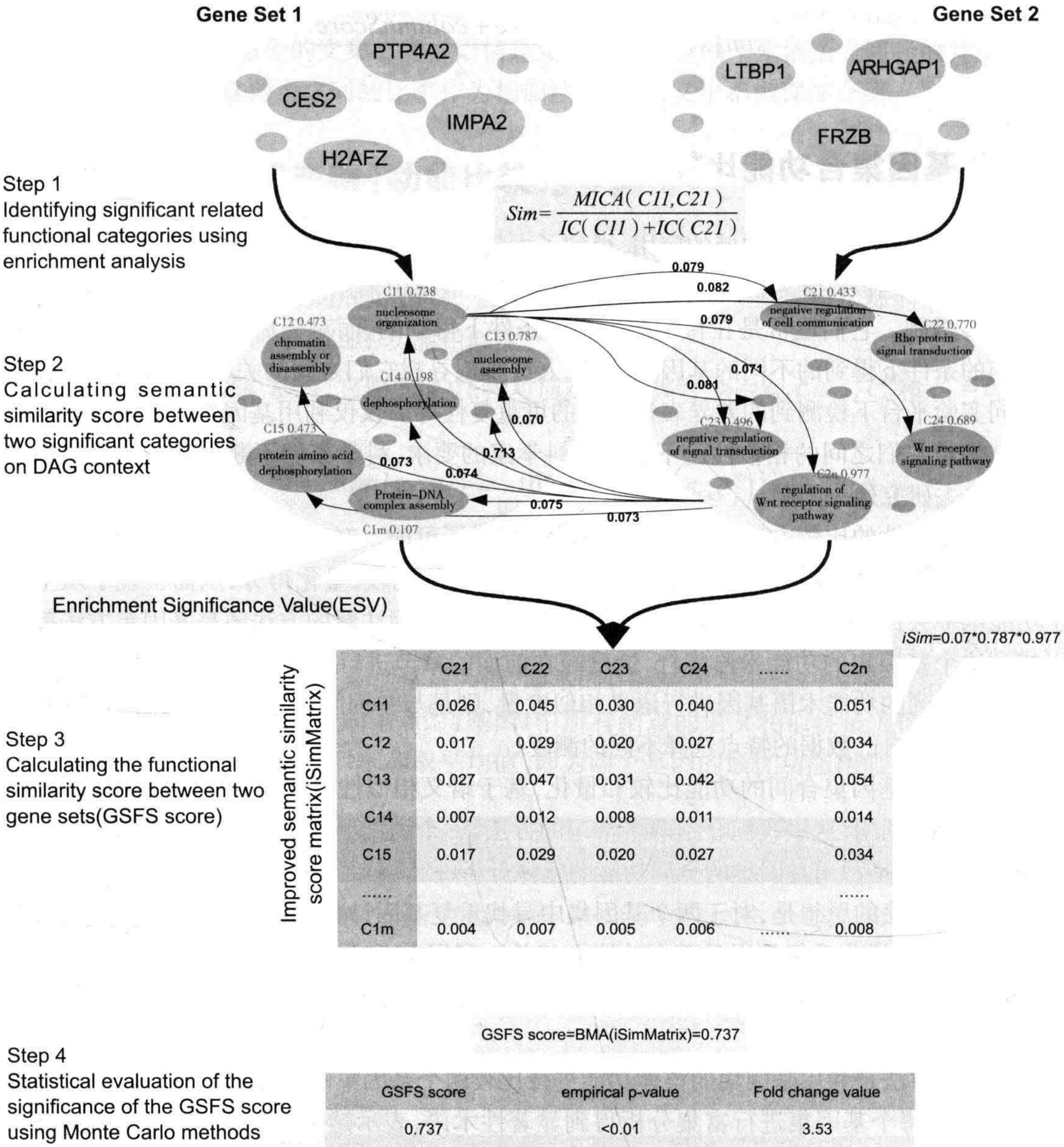


图3-21 利用基因集全局功能比较基因集功能

很多,这里列举了其中的一部分(表3-8):

表3-8 常用基于GO分析的语义相似性方法的平台及网址

数据库	网址
GOToolBox	http: //genome.crg.es/GOToolBox/
FunSimMat	http: //www.funsimmat.de/
FuSSiMeG	http: //xldb.fc.ul.pt/rebil/ssm/
G-SESAME	http: //bioinformatics.clemson.edu/G-SESAME/

续表

数据库	网址
GSFS	http: //bioinfo.hrbmu.edu.cn/GSFS
csbl.GO	R 包
GOSim	R 包
SemSim	R 包

（汪强虎 宁尚伟）

参考文献

1. Pesquita C, Faria D, Falcao A. O, et al. Semantic similarity in biomedical ontologies. *PLoS Comput Biol*, 2009, 5 : e1000443.
2. Resnik P. , Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. . *J. Artif. Intell. Res.* 1999, 11: 95–130.
3. Lv S, Li Y, Wang Q, et al. A novel method to quantify gene set functional association based on gene ontology. *J R Soc Interface*. 2011, 9 : 1063–72
4. Huang da W, Sherman B. T, Lempicki R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 2009, 37 : 1–13.
5. Huang da W, Sherman B. T, Lempicki R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009, 4 : 44–57.
6. Ashburner M, Ball C. A, Blake J. A, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000, 25 : 25–29.
7. Wixon J, Kell D. , The Kyoto encyclopedia of genes and genomes--KEGG. *Yeast*. 2000, 17, 48–55.
8. Wrzodek C, Drager A, Zell A. KEGGtranslator: visualizing and converting the KEGG PATHWAY database to various formats. *Bioinformatics*. 2011, 27 : 2314–2315.
9. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. *Mol Syst Biol*. 2007, 3 : 88.
10. Hu P, Bader G. , Wigle D. A, et al. , Computational prediction of cancer-gene function. *Nat Rev Cancer*. 2007, 7 : 23–34.
11. Wang Q, Sun J, Zhou M, et al. A novel network-based method for measuring the functional relationship between gene sets. *Bioinformatics*. 2011, 27 : 1521–1528.

第四章
CHAPTER 4

蛋白质结构分析
PROTEIN STRUCTURE ANALYSIS

蛋白质的各种各样的功能是以它们对与之相互作用分子的高度特异性为基础的，这就要求蛋白质具有相当的刚性空间结构。这些结构的微小改变常常会使蛋白质丧失活性或发生剧烈变化，从而使其功能发生改变甚至影响生理功能导致疾病的产生。有关蛋白质三维结构的知识是了解蛋白质如何行使功能所必需的。生物系统的高分辨结构信息将允许我们对生命系统的功能、对系统修饰或扰动的后果进行精确的解释和推理。这一结构信息的展现与日益增长的基因组、蛋白组、代谢组信息相联系，为分析生物医学问题提供了强大的研究背景。

第一节

蛋白质高级结构

Section 1 Advanced Structures of Protein

一、蛋白质的高级结构特征

分子生物学的中心法则确定了DNA与蛋白质氨基酸序列间的关系，称为第一套遗传密码子；确定蛋白质氨基酸序列与三维结构间的关系，被称之为“第二套遗传密码子”。蛋白质的一级结构(primary structure)就是蛋白质多肽链中氨基酸残基的排列顺序(sequence)，靠共价键维持多肽链的连接，而不涉及其空间排列，是蛋白质最基本的结构。它是由基因上遗传密码的排列顺序所决定的。各种氨基酸按遗传密码的顺序，通过肽键连接起来，成为多肽链，故肽键是蛋白质结构中的主键。蛋白质的一级结构决定了蛋白质的二级、三级等高级结构。成百亿的天然蛋白质各有其特殊的生物学活性，决定每一种蛋白质的生物学活性的结构特点，首先在于其肽链的氨基酸序列。由于组成蛋白质的20种氨基酸各具特殊的侧链，侧链基团的理化性质和空间排布各不相同，当它们按照不同的序列关系组合时，就可形成多种多样的空间结构和不同生物学活性的蛋白质分子。

(一) 蛋白质的二级结构

蛋白质二级结构(secondary structure)是指多肽链借助于氢键沿一维方向排列成具有周期性的结构的构象，是多肽链局部的空间结构(构象)，主要有α-螺旋、β-折叠、β-转角及无规卷曲等几种形式，它们是构成蛋白质高级结构的基本要素。

1. α-螺旋(α-helix)　α-螺旋是蛋白质中最常见最典型含量最丰富的二级结构元件。在α-螺旋中，与α碳原子相连的两个二面角都是恒定的，并且每圈螺旋包含3.6个氨基酸残基，残基侧链伸向外侧，同一肽链上的每个残基的酰胺氢和位于它后面的第4个残基上的羰基氧彼此之间形成氢键。这种氢键大致与螺旋轴平行。一条多肽链呈α-螺旋构象的推动力就是所有肽键上的酰胺氢和羰基氧之间形成的链内氢键。在水环境中，肽键上的酰胺氢和羰基氧既能形成内部(α-螺旋内)的氢键，也能与水分子形成氢键。典型的α-螺旋是由18个氨基酸残基形成的5圈螺旋，长约27Å。α-螺旋太长趋于形成纤维，不易形成球形。在大多数球状蛋白中，α-螺旋的平均长度约17Å，相当于11个氨基酸残基。

2. β-折叠(β-sheet)　β-折叠也是一种重复性的结构，可以看成是一种特殊的螺旋，是拉伸的α-螺旋，大多数球状蛋白质中，每股β-折叠链或β链(β-strand)的平均长度约

20 Å,相当于6.5个氨基酸残基,通常含有3~10个氨基酸残基。β-折叠链中同一肽段邻近肽键间很难形成氢键,只有通过较远距离的肽键之间形成氢键,将多股β-折叠链组合成一组β-折叠,一般称为β片层结构。通常分为平行式和反平行式两种类型,它们是通过肽链间或肽段间的氢键维系。构成β片层的几股β折叠链如果走向是相同的,则为平行的β片层;如果它们的走向是相反的,则是反平行的β片层。平行折叠片比反平行折叠片更规则且一般是大结构,反平行折叠片可以少到仅由两个β链组成,反平行的β折叠比平行的β折叠更为稳定。

3. β-转角(β-turn) β-转角是连接相同主链上α-螺旋、β-折叠等二级结构的关键结构。在β-转角中第一个残基的C=O与第四个残基的N-H氢键键合形成一个紧密的环,使β-转角成为比较稳定的结构,多处在蛋白质分子的表面,在这里改变多肽链方向的阻力比较小。β-转角可看成是由几个氨基酸残基构成的最小的反平行的β片层,即截短的发夹结构。β-转角的特定构象在一定程度上取决于它的组成氨基酸,某些氨基酸如脯氨酸和甘氨酸经常存在其中,由于甘氨酸缺少侧链(只有一个H),在β-转角中能很好地调整其他残基的空间阻碍,因此是立体化学上最合适的氨基酸;而脯氨酸具有换装结构和固定的角,因此在一定程度上迫使β-转角形成,促使多肽自身回折且这些回折有助于反平行β折叠片的形成。大多数β-转角存在于分子的表面,极少出现在分子的内部。β转角及其附近比整个分子有更大的亲水性。

β-凸起是一种小片的非重复结构,能单独存在,但大多数经常作为反平行β-折叠片中的一种不规则情况而存在。β-凸起可认为是β-折叠链中额外插入的一个残基,它使得在两个正常氢键之间在凸起折叠链上是两个残基,而另一侧的正常链上是一个残基。

4. Ω环形(Ω loop) Ω环形具有准有序结构,从形式上可以看成是β-转角的延伸,这类肽段的外形和希腊字母Ω相似,故被称为Ω环形。Ω环形的可变性比转角更大。在直接和蛋白质生物活性有关、有更大活动性的位点绝大多数是由转角和环形构成的。

5. 无规卷曲(random coil) 无规则卷曲或称卷曲(coil),泛指那些不能被归入明确的二级结构如折叠片或螺旋的多肽区段,是规律性较低而难以描述的特殊类型二级结构。其所涉及的残基数量差异大,整体外形变化大,可采取多种折叠形式,且不同构象间的能量差异小而容易相互转变,故其结构的规律性很低,但每一种蛋白质肽链中存在的这一类型"无规"肽段的空间构象是大致相同的。它们也像其他二级结构那样是明确而稳定的结构,否则蛋白质就不可能形成三维空间上每维都具周期性结构的晶体。它们受侧链相互作用的影响很大,经常构成酶活性部位和其他蛋白质特异的功能部位。无规卷曲在球状蛋白质表面出现较多,也是连接其他规则二级结构的结构模式。

(二) 超二级结构

超二级结构(supersecondary structure)指位于同一主链的多个二级结构组装形成的特定组装体,可直接作为三级结构的或结构域的组成单元,是从蛋白质二级结构形成三级结构的一个过渡结构形式,也称为立体结构形成的模体。α螺旋、β折叠和β转角的二级结构自身可形成超二级结构,不同的二级结构组合可以形成多种类型的超二级结构。

超二级结构主要有如下类型:①β-转角或Ω环等连接连续四个α-螺旋形成的四α-螺旋捆;②中部固定位置含有亮氨酸及其他疏水侧链氨基酸残基、在螺旋两端含有强亲水侧

链氨基酸的α-螺旋组成的亮氨酸拉链(leucine zipper); ③一条主链中相邻七个两亲α-螺旋通过过渡结构形成的七次穿膜螺旋组; ④连续主链中两段α-螺旋连接三段β-折叠链形成的Rossmann折叠; ⑤β-转角连接α-螺旋构成的α-螺旋-β-转角-α-螺旋; ⑥Ω环连接α-螺旋-α-螺旋-Ω环-α-螺旋等; ⑦β-折叠都为超二级结构。超二级结构通常并不对应生物化学功能,但其结构模式是解析蛋白质组装机制的关键信息之一。

结构域(domain)也是蛋白质构象中二级结构与三级结构之间的一个层次,它是在较大的蛋白质分子中,由于多肽链上相邻的超二级结构紧密联系,形成在空间上可以与蛋白质亚基结构明显区别的结构形态。一般每个结构域由约100~200个氨基酸残基组成,各有独特的空间构象,可承担特定的生物化学功能。

(三) 三级结构(tertiary structure)

蛋白质的一个引人注目的特征是它们都有确定的三维结构。一个伸展的或随机排布的多肽链没有任何生物活性,多肽链必须按照一定的规律折叠成三维结构,才具有生物活性。蛋白质三级结构即蛋白质分子中所有共价相连原子的空间相对位置,由多肽链在二级结构的基础上进一步盘绕和折叠形成;蛋白质如有特殊的必需辅基,其三级结构也包括来自这类辅基的原子的空间位置。稳定蛋白质三级结构主要靠氨基酸侧链之间的疏水相互作用、氢键、二硫键、范德华力和静电作用等。不同类型的蛋白质局部结构分解后可具有很高的相似性,但在三级结构层面不同蛋白质所体现的各自整体结构特征通常不同。

蛋白质按其"环境条件"的大体结构分类

1. 纤维状蛋白质　整条肽链几乎是单一的二级结构组成的巨大的、通常是缺水性的聚集体;其结构通常是高度氢键键合和高度规则的,且主要由不同肽链间的相互作用维系。在生物体内起到结构和支撑的作用。

2. 膜蛋白质　主要是指多次穿膜的膜蛋白,存在于缺水性的膜环境中,其膜内部分是高度规则的,也是高度氢键键合的,但大小上受限于膜的厚度。在膜内部分倾向于形成两亲的α-螺旋或β-折叠链;且形成疏水的在外侧,亲水的在内侧中间空心的圆桶状结构;内侧可作为亲水或极性物质的通道,连接这些膜内二级结构单元的肽段分布在膜的两侧,还承担其相应的生物功能。

3. 水溶性球状蛋白质　绝大多数的蛋白质的肽链折叠成为几乎球状的结构,存在于水中,较不规则(特别是小的球状蛋白质)。蛋白质的结构由其链内的相互作用维系,其中起重要作用的是在序列中远离但在空间上相邻的烃基(疏水)基团间的相互作用,有时还有肽链与辅因子的相互作用。一旦具有三级结构后,蛋白质内部变得更为紧密,其内部是大量的极性基团,而表面是以侧链的非极性残基为主导地位,极性和非极性残基这样的分布,使得肽链中大部分键的张角适合于稳定构象的形成。在蛋白质的内部存在有极少量的亲水的残基,而分子表面是一些疏水的残基,这些局部的构象具有相对偏高的能量,相对地处于较不稳定的状态,并以此行使蛋白质的功能。另外,蛋白质内部的部分水分子也和一些极性的基团或负电性的原子形成氢键,以此参与蛋白质功能的行使。

(四) 四级结构

四级结构是独立三级结构形成的复合物,其中每个独立三级结构为亚基(subunit),也称

为单体(monomer)。一些具有三级结构的肽链，通过其上面相互作用的基团，以特定的方式组装成为一种更高层次的结构，这个结构层次就是蛋白质的四级结构。其分类原则主要有：①按照亚基的数目分类：可分为寡聚和多聚两大类。在寡聚体中，亚基多数紧密地堆积而呈球形。多聚体可以成为线性结构。②按照亚基种类分类：可将蛋白质分为由相同亚基和不同亚基构成的两大类型。目前已知的蛋白质，绝大多数是由相同亚基构成的同源聚集体；由不同亚基组成的异源聚集体，也主要是2种或3种亚基组成的。③按照四级结构的外形分类：有的蛋白呈球形，另一些呈纤维状。亚基数目大于4的蛋白质，其四级结构可以呈现多种排列方式和不同的对称性。

具有四级结构的蛋白质通常有多个相同或不同的活性位点，比单纯的三级结构蛋白质具有更复杂的功能和调节机制。很多膜蛋白是由多个或多种亚基组成的具有四级结构的蛋白质，可以承担多种多样的功能，大多数是起通道和运转作用的蛋白质以及受体类蛋白质。

形成四级结构全部依靠非共价键相互作用，来自不同亚基的二级结构间可发生强的相互作用以稳定四级结构，如生成跨亚基的更大β-折叠结构或α-螺旋聚集体；其中，氢键、疏水相互作用和静电作用是主要维持力。为了形成稳定的四级结构，必然要求相互作用的任两个蛋白质之间的空间外形互补以增加接触面且理化性质互补。这些特征也是预测蛋白质间相互作用时有用的辅助判据。

从序列预测四级结构实际上是预测不同蛋白质间的相互作用，这是蛋白质功能预测的重要内容，也是结构生物信息学的重要任务。

二、蛋白质结构域与家族分类

蛋白质的复杂结构和功能依赖于多个结构域的协同；蛋白质缺失某个结构域(domain)则其必然缺失对应的生物化学功能。据蛋白质序列相似度或生物化学功能与结构的相似度可将蛋白质分类为家族(family)；同一家族蛋白质有某种类似的生物化学功能或者类似的高级结构。因此，了解蛋白质结构域及家族分类信息，对于蛋白质结构分析有着很重要的意义。

(一) 蛋白质结构域

结构域是构成蛋白质亚基的紧密球状区域，为介于二级与三级结构之间的一种结构层次；是蛋白质中可以具有独立三级结构的部分，通常由一个基因外显子编码，并可具有特定的功能。在较大的蛋白质中结构域之间通过较短的多肽柔性区互相连接；蛋白质的结构域有时还可分为一些次级结构，称为组件(module)。组件是在稳定的蛋白质功能域中常见的一种进化上保守而又独立的折叠单位，也是在进化压力下发生外显子迁移的基本单位，它还参与新基因的产生。结构域可以作为蛋白质三级结构的组件，通常不具有完整的生物学功能但有特殊的生物化学作用，这也是结构域与三级结构的关键区别。

一级结构氨基酸序列的某些区域相邻的氨基酸残基形成有规则的二级结构(如α-螺旋、β-折叠、β-转角和无规卷曲等)；然后再把相邻的二级结构片段集装在一起，形成超二级结构；在此基础上，多肽链再进一步折叠，成为近乎球状的三级结构就可成为一个结构域。最常见的结构域含有约100~200个氨基酸残基，一般至少40个，多的可达400个以上；对于较

小的蛋白质分子或亚基,其结构和功能都较简单,难以区分出独立的不同结构域,这类蛋白质属于单结构域分子(如卵溶菌酶等)。对于一个较大球状蛋白质分子来说,一条很长多肽链往往由两个或两个以上相对独立的三维实体缔合形成三维结构体。从功能角度看,很多蛋白质属于多结构域的蛋白,其功能位点基本都位于结构域之间,这是由于:①通过结构域容易构建具有特定三维排布的功能中心;②结构域之间常只有一段肽链相连,使结构域之间容易发生相对运动,这将有利于功能位点与对应成分相互作用或施加应力,有利于产生别构效应而对蛋白质的功能实现精细调节。

(二) 蛋白质家族分类

蛋白质结构域对于了解蛋白质的结构和功能意义重大。目前建立在结构域基础上的蛋白质家族数据库有PROSITE、PRINTS、Pfam、SMART、SWISS、PROT、ProDom 和BLOCKS等。因为每个数据库都有各自的分类原则和积分标准,将它们结合起来可以更准确地归类蛋白质家族和描绘结构域。随之出现了InterPro数据库,它是将蛋白质的结构域和功能位点加以统一而建立的数据库资源。InterPro联合PROSITE、PRINTS、Pfam和ProDom 四个独立完整的蛋白质结构域数据库组成站点,共包含18 349个条目,再现了5149个结构域、11 082个蛋白质家族等信息。此外, PDB、SCOP、CATH、HOMSTRAD、CAMPASS等蛋白质结构数据库运用不同的原理来识别结构相似的蛋白质超家族;蛋白质的结构域在进化过程中比序列保守,一些通过核苷酸序列识别不到的蛋白质超家族在这些数据库中可以被用户检索查询得到(表4-1)。

表4-1 常用的蛋白质结构域查询网址

数据库	网址
PROSITE	http: //www.expasy.ch/prosite/
BLOCKS	http: //blocks.fhcrc.org/
Pfam	http: //pfam.sanger.ac.uk/
ProDOM	http: //prodom.prabi.fr/
SMART	http: //smart.embl-heidelberg.de/
InterPro	http: //www.ebi.ac.uk/interpro/
SBASE	http: //www.icgeb.trieste.it/sbase
PRINT	http: //www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/PRINTS.html

三、蛋白质结构可视化软件

目前已有蛋白质高级结构数据存储的通用格式和数据库,可通过软件将蛋白质高级结构可视化,这些资源是蛋白质高级结构信息分析的关键基础之一。可视化分析蛋白质的高级结构有利于从原子间相互作用的层次理解生命活动过程的信息控制机制,理解蛋白质分子结构和各种微观性质与宏观性质之间的关系。

(一) 常用蛋白质分子图形系统

目前,蛋白质分子图形学软件已很普及;蛋白质结构数据可从蛋白质数据库中直接获

得。只要安装蛋白质分子图形学软件，并获得所需蛋白质结构数据，配以免费的小分子图形设计系统（如ACD FREE）或商业软件，就可开展结构生物信息学的探索性工作。

这里，着重介绍蛋白质三维图形相关的软件Pymol的基本应用。

1. 软件安装、启动和教程　Pymol可在http: //www.pymol.org/寻找链接下载，与其他Windows系统下软件的安装相同。

Pymol启动后显示双界面，对分子进行操作的常用命令及按钮都集成在一个图形显示界面，但文件读入、背景设置、操作转变、图像输出、特征分析等功能主要集中在另一个不显示分子图形且使用下拉菜单的界面，并带有命令行操作模式；关闭任意窗口则程序关闭。图形界面左上侧列出主要的可操作对象并分成几个层次，包括所选对象、蛋白质、整体等；每个层次的对象有五种主要操作：动作（A;action）、显示（S; Show）、隐藏（H; hide）、标记（L;Label）、上色（C;Color）。Dispaly下拉菜单中可设置背景（论文中这类图一般用白色背景，而报告中常用黑色背景以增加视觉效果），Wizard中有测定分子常用性质的模块，包括距离、电荷等，以及尝试进行蛋白质分子改造的功能。需要仔细阅读每个下拉菜单包含的功能才有利于发挥该软件的作用。可先读入教程文件进行学习（图4-1）。

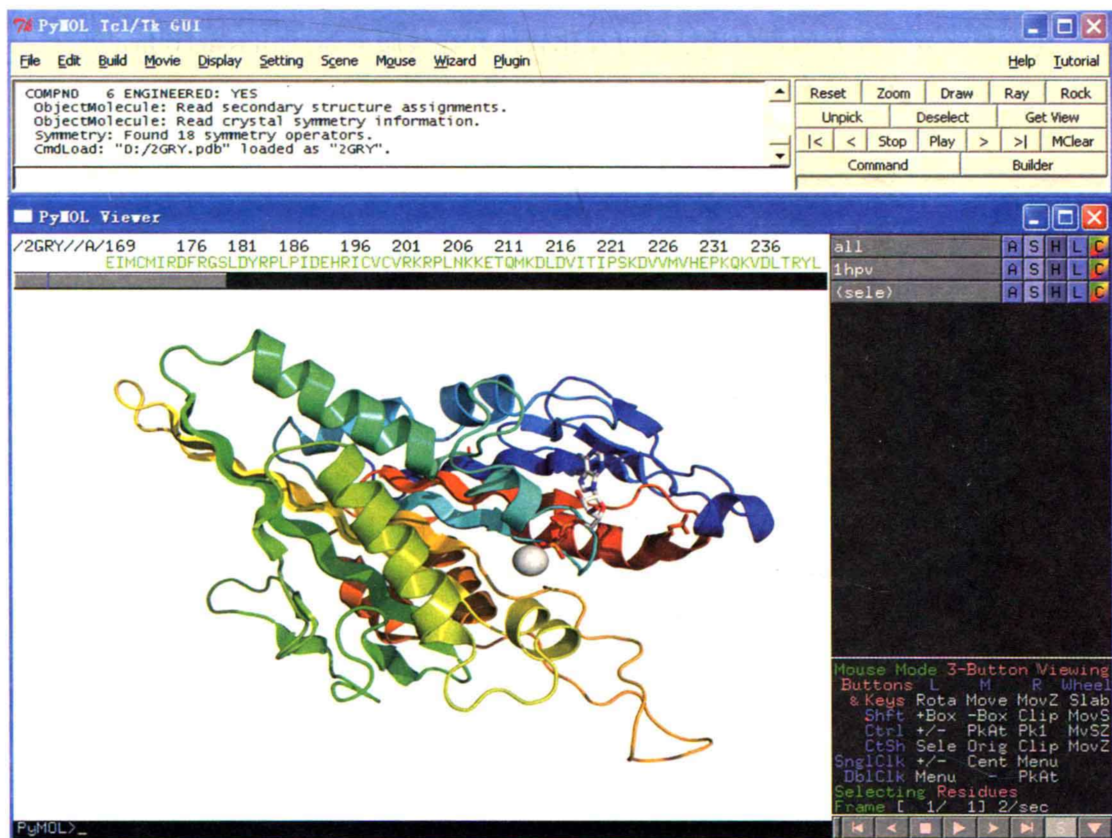


图4-1　Pymol启动后的两个操作界面（上下两个窗口），随后读入教程所用结构

2. 主要的分子图形操作和性质测定　鼠标是主要的图形操作工具，左键旋转图形，右键调整大小，也可在另一个窗口的下拉菜单中选择放大缩小；可设置鼠标的模式（两键与三键鼠标等，见Mouse下拉菜单）。可显示蛋白质中每条肽链的序列和非蛋白质成分（单击图形界

面右下角字母S或在Display下拉菜单中选择）；鼠标左键单击序列选中特定待操作的残基可同时显示对象所在位置，在Wizard中有多种性质测定功能，可灵活使用。

Pymol是强大的分子图形显示和基本特征测定系统，在带有专业显卡的计算机上输出图形更绚丽。但Pymol对非英文文件名和长文件名支持不够。Pymol自带二级结构定义词典但对α-螺旋的定义不严格，有时会给出一些不尽合理的α-螺旋；不过有些商业软件不能识别同源建模所得蛋白质中的二级结构而Pymol可以识别这些二级结构，这是Pymol的一个优势。

（二）集成的分子模拟与分析的图形学系统

集成结构生物信息学、分子操作绝大部分功能和MD模拟轨迹分析等功能的商业软件已面市，如InsightⅡ、Discover Studio和Sybyl等；这些商业图形操作界面系统价格不菲，但可集成在图形界面进行分子模拟、分子对接和分子改造等操作，并有各种高质量的图形显示，对应用研究人员无疑可事半功倍。

（三）其他的蛋白质可视化软件介绍

还有很多界面友好的蛋白质结构可视化软件和在线服务器，如RasMol和Jmol等，已与PDB数据库链接；另外还有Cn3D、Mage、KiNG等可视化软件（表4-2）。

表4-2 目前常用的蛋白质可视化软件

软件名称	主要功能	下载地址
RasMol	直观再现生物分子3D微观立体结构；提供可以旋转等多个模式效果图；提供多种结果图片存储形式；提供命令行操作，源代码开放用户可自行维护	http://www.bernstein-plus-sons.com/software/rasmol/
Jmol	以3D形式查看蛋白质等生物大分子化学结构，提供命令行操作，提供结构查询工具，基于网络界面可通过网址或本地文件读取结构，无需安装（Jmol提供的功能适用于小分子，晶体，材料和生物分子）	http://jmol.sourceforge.net/
Cn3D	生物分子三维结构、序列以及序列比对结果的可视化工具；读取输入数据格式为MMDB格式文件，不能读取PDB格式文件；可紧密联系结构与序列信息，可根据基于结构的序列比较显示分子结构之间的关系；可自定义标签特征，输出结果格式多样，并可对结果进行文献注释；通过网络浏览器来作为NCBI的Entrez系统的一个辅助工具，也可作为一个独立的程序使用	http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml
QuickPDB	用JAVA编译的显示结构和序列的工具；网络浏览可直接显示序列信息，可以控制设置残基属性等；支持多种文件格式输入、可以不同形式显示三维结构	http://www.sdsc.edu/pb/Software.html
Mage	广泛应用于教学与研究中，输入为*.kinemage文件格式，该文件内含有蛋白质结构的各种信息与相关命令；可实时旋转效果图、并对效果图进行蛋白质结构的三维动画演示，部分图像可隐藏和显示；输出格式为.kinemage，也可以多种其他格式输出	http://kinemage.biochem.duke.edu/software/mage.php

续表

软件名称	主要功能	下载地址
VMD	主要处理目标是分子动力学数据,可对生物分子进行结构分析和表征;提供多用途交互式图形界面操作;开源并提供强大脚本语言,可用于程序扩展	http://www.ks.uiuc.edu/Research/vmd/
KiNG	KiNG即Kinemage,是在Mage,JavaMage和Kinemage软件基础上发展起来的三维分子显示软件,可展示生物大分子结构	http://kinemage.biochem.duke.edu/software/king.php
Spdbv	即Swiss-Pdb Viewer或DeepViewer。可同时分析几个蛋白质的PDB文件并分析结构相似性、比较活性位点或其他有关位点;可以很容易获得氢键、角度、原子距离、氨基酸突变等数据;可直接从软件连接到Swiss-Model服务器对蛋白质理论立体结构进行构建,并调用POV-Ray软件生成高质量的结构图像	http://mac.softpedia.com/get/Math-Scientific/SPDBV.shtml
WebMol	用JAVA语言编译的结构呈现程序,网络浏览,可从URL上载结构	http://www.cmpharm.ucsf.edu/~walther/webmol/download.html
Raster3D	可显示蛋白质三维结构并生成蛋白质结构的分子艺术图片(TIFF格式与JPG格式)	http://www.fyxm.net/Raster3D-93918.html

第二节

蛋白质的结构解析与预测

Section 2 Analysis and Prediction of Protein Structure

一、蛋白质结构实验检测技术与结构解析

蛋白质及其复合物、组装体的完整精确的三维结构的测定是研究生命活动中分子结构和功能关系，揭示生命现象本质的基础。根据蛋白质的状态，测定蛋白质三维结构的方法分为两大类：①应用X射线晶体衍射图谱法、冷冻电子显微镜技术和中子衍射法测定晶体中的蛋白质分子构象；②应用磁共振波谱分析法、园（圆）二色性光谱法、激光拉曼光谱法、荧光光谱法、紫外差光谱法和氢放射性核素交换法等测定溶液中的蛋白质构象。近几年来近场光学光谱技术、表面等离子体激元共振技术、化学交联法等也用于获得蛋白质的静态或动态结构信息。

（一）蛋白质晶体结构X-衍射分析

X-射线晶体分析法（X-ray diffraction crystallography）是解析生物大分子结构的基本方法，也是目前分辨率最高的方法，已用于大量蛋白质的三维结构的解析。该法需要将待分析的蛋白质形成晶体，所用蛋白质样品量很大，故常将该蛋白的基因克隆到表达载体，在特定宿主细胞（如大肠埃希菌）中诱导表达，纯化后优化条件结晶；然后将晶体进行X射线衍射，收集并整合相应的衍射图谱，通过复杂的计算和数据解析过程得到蛋白质中的原子坐标信息。

高通量晶体结构解析主要涉及数据处理与分析、重原子的定位、密度修饰、分子替换、图形整合、模型加工和确认等环节。X 射线衍射实验记录的是衍射点的强度和方位，从衍射点的强度可推算出该点的结构振幅，而该衍射点的相角信息却无法从实验中直接得到。因此，晶体结构分析的核心问题就是要找出各衍射点的位相。晶体衍射数据分析的常用软件有XRayView、SOLVE和RESOLVE等。XRayView 适用于X射线衍射晶体数据的交互式动态分析，涉及晶胞的构建、晶格的确定、系统消光、旋转摄影、空间群的确定及Laue群对称性等。X-PLOR是适用于计算结构生物学的程序系统，通过经验能量函数及实验数据的限定，进行大分子空间构象的开发，该程序主要用于对X-射线衍射数据及NMR核磁共振数据的分析。优化蛋白质结晶条件、快速处理晶体衍射数据是目前蛋白质晶体结构分析的两大难题；发展高通量的蛋白结晶技术和高可靠性的结构解析技术，是当前结构生物学的重要任务。随着晶体结构的运算法则和计算机科学的发展，新一代的自动化分析软件将进一步解决高通量

结构分析的技术问题，并将适时处理各种衍射数据和加快图形整合过程。

(二) 磁共振波谱分析

磁共振（nuclear magnetic resonance，NMR）是以组成蛋白质的最基本元素：C、N、H、O等原子的原子核为探针检测蛋白质的结构信息的。而原子核自旋量子数I=1/2的核：1H、13C、15N 是多维磁共振检测的主要对象。在外加静磁场H_0，即恒定超导磁场的作用下，这些核自旋量子数不为零的原子核会发生能级分裂。如果同时将射频磁场H_1作用到原子核系统上，当射频场频率ω满足关系式：$\omega = \gamma H_0$，原子核将吸收射频场能量从低能级跃迁到高能级。这种共振跃迁现象就是磁共振现象。上述关系式即为磁共振条件。式中γ为旋磁比。不同的原子核的旋磁比不同，因而1H，13C，15N 的磁共振频率不同，所以有磁共振氢谱、碳谱、氮谱之分。对蛋白质溶液样品进行各种类型的同核或异核多维磁共振实验，并由这些实验所提供的磁共振波谱信息，建立用于溶液中蛋白质三维结构计算的磁共振数据文件，这是多维磁共振方法确定溶液中蛋白质三维结构的思想。

目前，用NMR测定蛋白质结构的数据处理涉及许多复杂的算法。磁共振波谱的谱峰包含有相当丰富的与蛋白质分子结构有关的波谱信息，它们由波谱参数表示。此过程中首先是将磁共振的信号经过傅里叶变换转换为不同的峰值，然后采集各种不同的峰组成图谱，并筛选出具有特定结构特征的图谱。这些过程常用NMRPipe和SPARKY软件（http://www.cgl.ucsf.edu/home/sparky/）处理，也使用DGⅡ、XEASY、DYANA和GARANT等软件分析计算蛋白质三维结构、侧链或骨架结构。即在具体计算蛋白质结构过程中，无论是运用哪一个基于距离几何方法的计算软件，都是将磁共振波谱提供的NOE 和J 耦合常数数据转换为用于结构计算的距离约束、二面角约束、手性等结构数据文件，其中也包括形成氢键的原子对之间距离的约束；结合从蛋白质氨基酸组分得到的蛋白质分子中的键角、键长、手性等经验数据，建立约束矩阵。然后，将距离空间的约束矩阵转换为坐标空间的矩阵。接着，由坐标空间矩阵构建蛋白质三维结构的初始结构模型。最后，运用模拟退火等计算方法对初始结构进行优化，并由分子动力学进行能量最小化计算，由此得到一组收敛的蛋白质三维结构的空间坐标，即获得由磁共振实验数据导出的蛋白质溶液三维结构的一系列可能的构象集合。

与X-衍射晶体分析技术相比较，NMR技术尽管在蛋白质结构测定中限制较大，但其无需制备晶体，故NMR法常用于解析无法获得晶体的蛋白质或膜蛋白的结构。目前，NMR技术主要用于解析分子量在20kD以下且水溶性很好但培养晶体困难的蛋白质结构。由于其分析过程可在溶液状态进行，从而得到蛋白质分子在溶液中的构象，条件更接近于蛋白质的生理状态，是研究蛋白质的折叠和构象稳定性对生理环境温度、盐浓度、pH等环境条件变化敏感性的重要工具。在溶液环境中，可以观察到整个结构表面的一些松散肽链的运动性，而蛋白质的功能部位往往是在整个结构的表面，因此，NMR是研究蛋白质与蛋白质、蛋白质与小分子配体间相互作用的动力学特征和性质的有效手段。随着NMR技术的发展，NMR所用磁场强度的增强、计算资源的提升和分析软件的进一步发展完善，磁共振技术在蛋白质结构解析领域的应用会越来越广泛。

(三) 冷冻电子显微镜技术

冷冻电子显微镜（cryoelectron microscopy）技术已成为研究生物大分子结构与功能的强

有力手段。该技术大致包括样品制备、数据采集和图像处理以及三维重构等环节，其确定三维结构的方法主要有电子晶体学方法、单粒子重构法和电子断层成像技术。这种方法采用高压快速液氮冷冻方法使样品包埋在玻璃态的水环境中，这种环境也接近于生理状态，减少了样品在制备过程中的结构破坏，以便观察生物大分子在天然状态下的结构；同时冷冻的速度极快，这就有可能把细胞在其生理活动（例如，肌肉收缩）的某些特定时刻固定下来，并显示此时的结构特点，进而可通过不同功能状态的瞬时构象变化来研究生物分子的功能。故冷冻电镜获得的是处于天然状态下未经染色的分子的二维投影像。将样品进行不同角度倾斜所获得的数据进行综合分析，并依据样品的不同特性使用不同的重构技术获得分子的结构，在此基础上观察多种成分的图像变化，可追踪生物大分子的装配及其动力学过程。

冷冻电子显微镜技术主要用于蛋白质及其复合物的外部形貌观察，可用不同的方法对均一的（如膜蛋白的二维晶体，二十面体对称的病毒等对称结构）和不均一的（如核糖体等）样品进行三维结构重构，同时可应用的蛋白质分子大小范围很宽。使用冷冻电子显微镜技术观察生物大分子的空间构象需要借助生物信息学方法和模式识别（pattern recognition）、数据库分析、同源建模（homology modeling）等技术的整合。由冷冻电镜技术所获得的蛋白质三维结构与X射线晶体技术得到的结构非常相似，而且其信噪比非常低，并且适合于膜蛋白的分析。可用于处理和分析数据的软件有CCP4、CNS、EM3D、Bsoft、EMStudio、IMAGIC等。此技术目前应用面并不太广，也没有形成相应的数据库。各种相关技术的发展和整合将为研究生命现象与本质提供强有力的技术手段。

（四）化学交联（cross-linking）法

近年来出现了一种可以获得蛋白质结构信息以及蛋白质相互作用信息的新方法—化学交联（cross-linking）法，即在蛋白质样品中加入适量的化学交联剂（chemical cross-linker），使蛋白质内部或不同蛋白质之间发生交联反应，实现对蛋白质中各个氨基酸侧链或官能团空间位置的定位，再应用现代质谱法鉴定氨基酸侧链或官能团，获得氨基酸或官能团的相对空间距离，构建蛋白质的空间结构及蛋白质复合体亚基的空间排列位置，以及获得蛋白质相互作用的信息。通常采用MS2Assign 和MS2PRO软件对交联肽段的MS/MS 进行分析；SearchXLinks 软件可用于蛋白质二硫键质谱分析；利用general protein/mass analysis for windows（GPAMW）软件对MALDI-TOFMS 或ESI-MS 对水解后的混合肽段进行分析；还有其他的软件可用于化学交联质谱数据分析如Automated Spectrum Assignment Program（ASPA）VirtualMSLab等。化学交联法与质谱法的有机结合有如下优点：①被分析蛋白质或蛋白质复合体的分子量（molecular weight，MW）从理论上说是无限的，如果被分析物的分子量过大，可以采用Bottom-up分析策略；②质谱分析速度十分迅速，可以极大缩短分析时间；③质谱检测灵敏度极高，被分析的蛋白质只需飞摩尔（femtomole）级的量就可满足实验要求；④易于获得溶液中蛋白质三维结构信息，并可鉴定出蛋白质的可变结构域；⑤化学交联法结合质谱法在膜蛋白研究领域的应用与其他传统方法相比有其独特的优势。

二、蛋白质结构数据库

蛋白质三维结构数据库是一类重要的生物分子信息数据库，是结构生物信息学的关键

组成。随着X-射线晶体衍射技术、NMR和冷冻电子显微镜技术等的发展，很多蛋白质的结构已被测定；随着蛋白质结构分类研究的深入，出现了蛋白质家族、折叠模式、结构域、回环等数据库。总体而言，目前常用的蛋白质结构数据库主要是存储蛋白质结构的PDB数据库（protein data bank，PDB）、进行蛋白质结构比较的SCOP和CATH，及存储次级结构的targetDB、FSSP、DSSP等。

（一）蛋白质三维结构数据库PBD

PDB是用于保存生物大分子结构数据的常用数据库，由美国Brookhaven国家实验室于1971年创建。1998年10月为适应结构基因组和生物信息学研究的需要，由美国国家科学基金委员会、能源部和卫生研究院资助成立了结构生物学合作研究协会（research collaboratory for structural bioinformat-ics，RCSB）。之后PDB数据库的维护主要是由该组织负责，目前主要成员为拉特格斯大学（Rutgers University）、圣地亚哥超级计算中心（San Diego supercomputer center，SDSC）和国家标准化研究所（national institutes of standards and technology，NIST）。PDB数据库网站主页见图4-2。

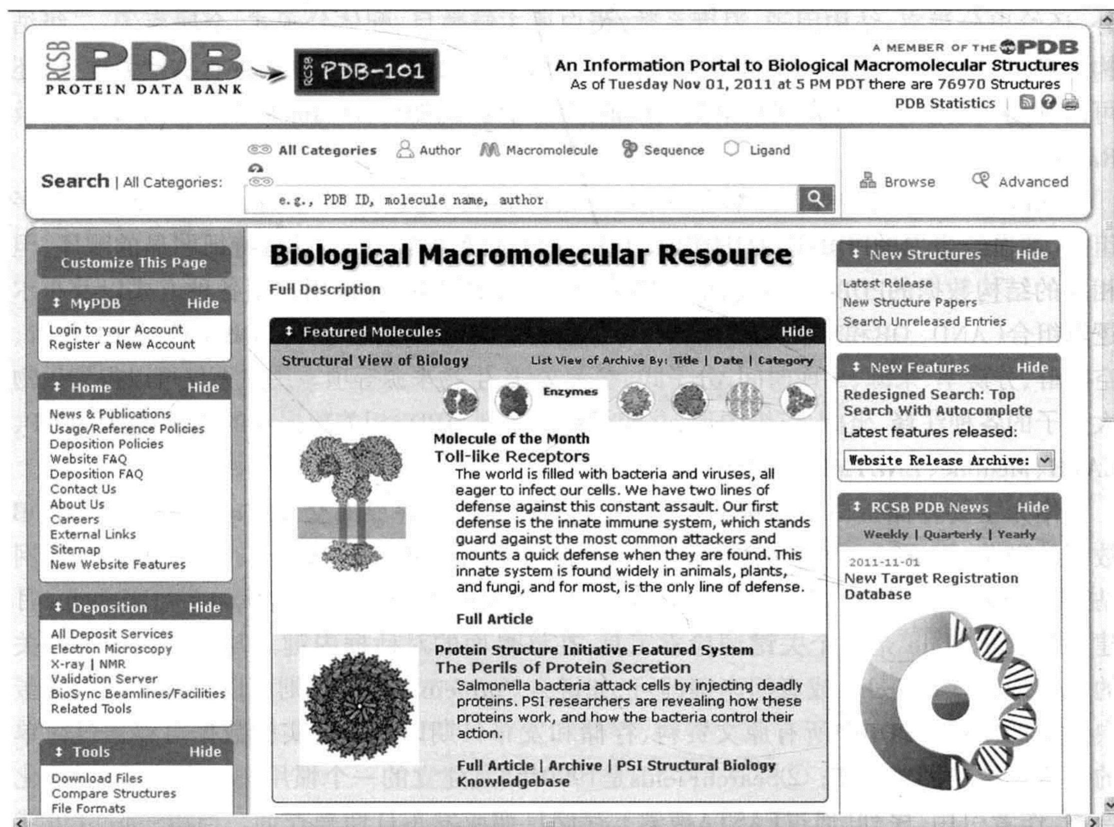


图4-2　PDB（Protein Data Bank）数据库网站主页

PDB中包含了通过X射线单晶衍射、磁共振、电子衍射等实验手段确定的蛋白质、多糖、核酸等生物大分子的三维结构数据。目前PDB数据库的信息每周进行更新，截止到2011年11月1日，PDB总共收录了76 970条结构数据，其中，收录蛋白质结构为71 309条，收录核酸

3326条。详细数据见表4-3。

表4-3 PDB数据库收录条目一览表

实验方法	分子类型				总数
	蛋白质	核酸	蛋白/核酸复合物	其他	
X-射线衍射	62 894	1323	3053	2	67 272
NMR	7970	960	179	7	9116
电镜	262	22	97	0	381
其他	133	4	5	13	155
总数	71 300	2312	3335	23	76 970

PDB数据库以文本文件的方式存放数据，每个分子各用一个独立的文件存放。文件中除了原子坐标外，还包括物种来源、化合物名称、结构以及有关文献等基本注释信息。此外，还给出分辨率、结构因子、温度系数、蛋白质主链数目、配体分子式、金属离子、二级结构信息、二硫键位置等和结构有关的数据。除了能以文本编辑的方式查看这些数据外，还可以利用一些图形软件直观观察蛋白质的三维结构，例如VMD、Jmol、Swiss-PDBviewer及RasMol等。

在PDB中收集的结构数据都有一个唯一的PDB-ID，它包含4个字符，由大写字母和数字组成（如血红蛋白的PDB-ID为4HHB）。PDB-ID编码系统较复杂，没有特征明显的顺序，但相关的结构数据的PDB-ID仍然有明显的联系。PDB数据库允许用户用各种方式以及布尔逻辑组合（AND、OR和NOT）进行检索，可检索的字段包括功能类别、PDB代码、名称、作者、空间群、分辨率、来源、入库时间、分子式、参考文献、生物来源等项。用户不仅可以得到生物大分子的各种注释、坐标、三维图形，并能得到一系列与PDB相关数据库的链接，包括SCOP、CATH、Medline、ENZYME、SWISS-3DIMAGE等。

作为主要存储蛋白质结构的数据库，PDB还提供多种界面交互方式实现用户对PDB数据的浏览，可通过三种查询方式对其主要服务器站点SDSC、Rutgers、NIST 和其镜像网站进行查询，也可进行相应数据的下载操作。数据库的查询方式（表4-4）：①1999年2月建立的SearchLite 是一个关键词检索工具，在该界面的对话框内键入与生物大分子相关的关键词，点“Search”或者回车键即可，如键入“protein kinase”，则可以查询所有包含蛋白激酶的结构。PDB中所有原文资料、存储和发布日期以及一些实验数据可以通过简单的浏览或结构浏览得到；②SearchFields是1999年5月建立的一个惯用浏览方式，可以用化合物、作者引用、序列（通过FASTA搜索）、存储日期或发布日期来查询。当用SearchLite或SearchFields浏览时，在“Query Result Brower”的界面可得到一些综合信息及图表中的详细信息，并可下载PDB中系列数据文件，下载的数据以纯文本格式或压缩文件的形式保存。“Struture Explorer”界面提供每个蛋白质结构的信息以及与许多大分子结构数据库的交叉链接。

表4-4　数据库查询方式

查询方式	使用方法
SearchLite	PDB所包含的任意词或词组
SearchFields	1.一般信息: PDB编码,作者以用,链型(蛋白质、DNA等), PDB HEADER,试验方法,存储或发布日期,复合物资料, BC数字或上下文检索 2. 序列或二级结构: 链长, FASTA 检索,短序列方式和二级结构内容检索 3. 晶体试验信息: 溶剂,空间基团,单体相关参数
Status	PDB编码,存储信息作者,题目,存储日期或发布日期

【例4-1】在PDB数据库中检索人类驱动蛋白相关的结构信息和可视化过程(图4-3)。

利用搜索关键字"HUMAN KINESIN"在PDB数据库主页搜索框内进行搜索;点击查

第一步:
输入关键字"HUMAN KINESIN"

第二步:
选择人体泛肽激酶马达结构域1BG2

图4-3　在PDB数据库中搜索人类驱动蛋白结构的结果

询结果页面的一个检索条目1BG2，打开其链接页面；在结果页面右侧列表信息中查看生物结构信息面板“Biological Assembly”部分；点击“Biological Assembly”面板查看1BG2结构图（需要JavaScript插件）；在结果页面中，可查看提供该蛋白质结构的作者信息（Deposition Summary）及实验细节信息（Experimental Details，包括分辨率resolution、空间群space group和近体的单位晶胞尺度unit cell dimension等）。另外，还可以链接到其他一些浏览结构信息的可视化工具如Jmol、Kiosk等进行精细结构的观察和分析。

（二）蛋白质结构分类数据库SCOP

蛋白质结构分类数据库（structural classification of protein，SCOP）是对已知结构蛋白质进行分类的数据库（图4-4），根据不同蛋白的氨基酸组成及三级结构的相似性，详细描述已知

结构蛋白间的功能及进化关系；SCOP数据库的构建除了使用计算机程序外，主要依赖于人工验证。SCOP数据库建立于1994年，数据库中信息主要由Alexdi G Murzin和其同事每年更新。

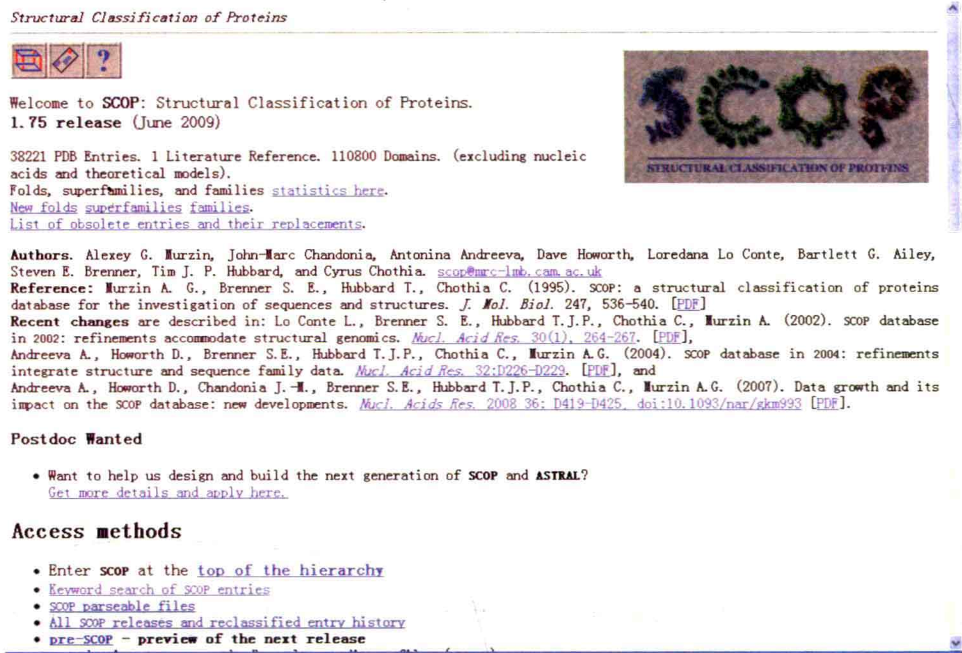


图4-4　SCOP数据库主页

目前SCOP数据库的最新版本是2009年6月发布的1.75版，在该版本中共含有38 221个已有结构的蛋白以及110 800个蛋白结构域，表4-5为SCOP数据库最新版本中详细的信息统计。

表4-5　SCOP数据库中1.75 版本中详细信息

蛋白质种类(Class)	折叠子的数目(Folds)	超家族的数目(Superfamilies)	家族的数目(Families)
全α螺旋蛋白	284	507	871
全β折叠蛋白	174	354	742
α螺旋和β折叠	147	244	803
α螺旋加β折叠	376	552	1055
复合结构域蛋白	66	66	89
膜蛋白	58	110	123
小蛋白	90	129	219
总和	1195	1962	3902

在SCOP数据库中，按照从简单到复杂的顺序对蛋白进行分类，分类基于四个层次，位于分类层次顶部的是类(class)，之后依次为家族(family)，超家族(super family)、折叠子(fold)、蛋白质结构域(protein domain)、单个PDB蛋白结构记录。SCOP数据库可以通过其分级结构导航进行浏览，用关键字、PDB标志码查询，或通过一个蛋白质序列进行同源搜索。在各个分类层次中，家族用来描述相近的蛋白质进化关系；超家族用来描述远源的进化关系；折

叠子用来描述空间的几何关系。在SCOP数据库中结构域又被分为以下几类：全α螺旋，全β折叠，α螺旋和β折叠，α螺旋加β折叠以及复合结构域。除此之外，SCOP提供一个非冗余的ASTRAIL序列库，这个库通常被用来评估各种序列比对算法；同时SCOP还提供一个PDB-ISL中介序列库，通过与这个库中序列的两两比对，可找到与未知结构序列远源的已知结构序列。除了显示蛋白质结构与进化的信息外，SCOP数据库通常可以链接到PDB、SP3D、NCBI Entrez等数据库来显示原子坐标，蛋白序列及同源蛋白信息。SCOP对多方用户都具有广泛的用途，全世界不同地区具有其相应的镜像站点。探究与所研究的蛋白质相近的结构空间区域时，蛋白质的分类层次有助于对蛋白质进行定位，而且数据库提供的交叉链接，方便对预测结果进行生物学解释。

(三) 蛋白质分类数据库CATH

另一个代表性蛋白质结构分类数据库是由伦敦大学于1993年开发和维护的CATH(图4-5)。该数据库的名称CATH分别是数据库中四种分类类别的首字母，即蛋白的种类(class, C)；蛋白中二级结构的构架(architecture, A)；蛋白的拓扑结构(topology, T)；蛋白质同源超家族(homologous superfamily, H)。SCOP注重从蛋白质进化角度进行分类，而CATH偏重于从结构角度对蛋白分类，同时数据库对蛋白进行分类时既使用计算机程序，也进行人工检查。

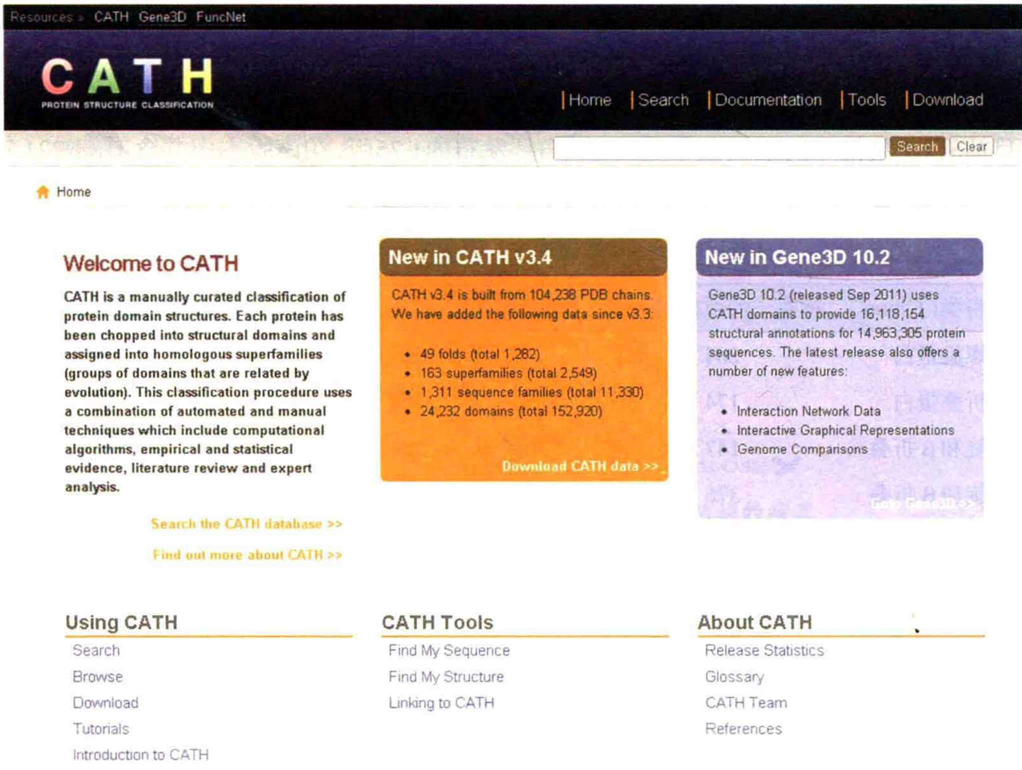


图4-5　CATH数据库主页

目前CATH数据库最新版本是2010年发布的3.4版，该版本中含有152 920个蛋白结构域，40个二级结构构架，1282个拓扑结构以及2549个同源蛋白质超家族。同PDB 蛋白结构数据库相似，每一个蛋白都会有一个不重复的标号，在CATH数据库中表现为不同分类层次都有

CATH号，并且不同水平的CATH号的标准不同。例如：位于CATH数据库最底层分类的蛋白种类类别，它的CATH号的范围为1~4，每一类之间的间隔量为1；而其他类别之间的间隔量为10。图4-6是CATH数据库中各个类别的层次划分结构。与SCOP不同的是，CATH把蛋白质分为4类，即全α、全β，α-β（α/β型和α+β型）和低二级结构类。低二级结构类是指二级结构成分含量很低的蛋白质分子。CATH数据库的第二个层次为由α螺旋和β折叠形成的超二级结构排列方式，而不考虑它们之间的连接关系。形象地说，就是蛋白质分子的构架，如同建筑物的立柱、横梁等主要部件，这一层次的分类主要依靠人工方法。第三个层次为拓扑结构，即二级结构的形状和二级结构间的联系。第四个层次为结构的同源性，它是先通过序列比较然后再用结构比较来确定的。除了以上提到的四种分类外，CATH数据库还有另外一种分类层次为序列层次，在这一层次上，只要结构域中的序列同源性大于35%，就被认为具有高度的结构和功能的相似性，从而被划分为在同一序列家族（sequence family）中；对于较大的结构域，则至少要有60%与小的结构域相同。

主要包括：α，α和β，β，“TIM折叠桶”结构，“三明治”结构，“肉冻卷”结构，黄素蛋白，β-内酰胺酶结构CATH数据库可以通过英国伦敦大学的生物分子结构和模拟实验室的网络服务器来实现用户数据的查询和分析。在CATH首页右上角的搜索框内输入待

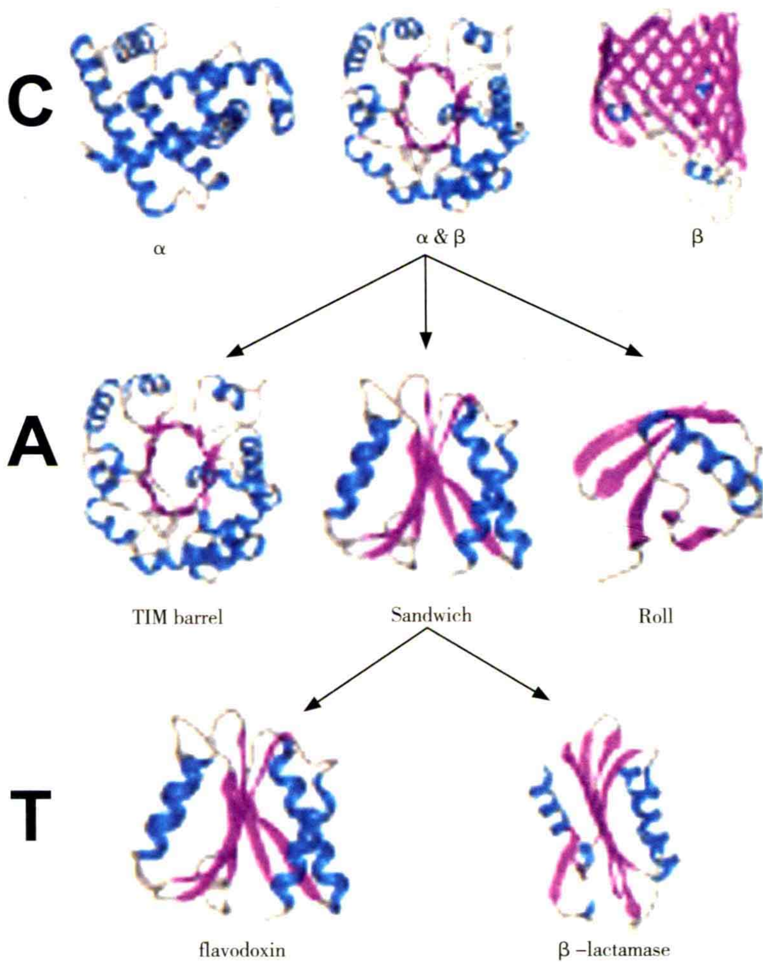


图4-6　CATH数据库中不同蛋白分类的结构图

查询关键字，点击“Quick Search”查询。CATH给用户提供了满足不同需求的数据查询方式，具体包括：

（1）搜索一个特定结构域信息，需要链接至“PDB code/Domain ID search”。用户搜索条目可以为CATH domain ID，CATH Chain ID或者PDB code，输入搜索条目关键字，点击首页右上角的“Quick Search”或者转到“Search CATH by ID/sequence/text”页面，利用“Search by ID/Keywords”模块进行搜索。

（2）搜索与用户给定结构或功能关键字相关的信息，需要链接至“Text Search”实现文本搜索查询。用户输入的搜索关键字可以是描述功能起源的“chaperone”或结构相关的“helix”。将搜索关键字输入到搜索框，点击首页右上角的“Quick Search”按钮进行查询或者转到“Search CATH by ID/sequence/text”页面，利用“Search by ID/Keywords”模块进行搜索。

（3）搜索CATH不同层次结构相关的信息，需要链接至“Browse the CATH hierarchy”，可查看数据库数据分类信息。也可通过“Search CATH by ID/sequence/text”页面，点击“Browse”按钮链接至“CATH hierarchy”（图4-7）。



图4-7 CATH数据库分类查询主页

其次，CATH数据库还提供了分析模块，可以提交感兴趣的查询条目，CATH数据库将为该查询条目提供相关的详细的结构和相应的功能信息。

在CATH数据库主页上，选择“Tools”，进入“CATHEDRAL Server”分析服务器，允许用户根据PDB ID标识或CATH code编码，进行相应的结构和功能分析。如用户可以从CATH数据库中获悉给定蛋白质FtsA（pdbid: 1e4f）在不同物种中的进化相关性以及与之密切相关的生物学功能信息。首先，检索CATHEDRAL Server服务器获取该蛋白质上所有的结构域、结构域家族和蛋白质超家族信息。需要指出的是，对于一个序列已知、结构未知的蛋白质，CATH数据库可以根据其结构比较算法将感兴趣的蛋白质与CATH数据库中的背景蛋白质进行相似结构搜索，最终确定出该蛋白质的结构和相应的结构域信息。然后，根据蛋白质FtaA

所有结构域所在的超家族信息，用户可以获悉该蛋白质处于核酸转移酶结构域家族（CATH code: 3.30.420.40）中。其次，利用CATH和Gene3D，检索CATH超家族3.30.420.40所包含的所有结构域信息。在CATH主页上输入结构域编码（1e4f+结构域标识/链标识），检索获取相应的结构域信息；在H-level同源超家族层，可以找到3.30.420.40，点击进入可以查看该同源超家族中其他已知结构域的结构信息；通过结构域链接，可以获取其结构信息和特定的分子细胞的功能信息等。具体内容可登录到数据库站点查看，这里不作赘述。

（四）其他常用蛋白质结构数据库

1. SWISS-MODEL数据库　SWISS-MODEL数据库收录的蛋白质结构都是使用SWISS-MODEL同源建模（homology-modelling）对Swiss-Prot蛋白序列数据库或其他蛋白质序列进行自动同源建模所得到的结构数据，该数据库保持定期更新。建立该数据库的主要目的在于提供最新的蛋白质3D结构注释信息。

至2011年6月，SWISS-MODEL数据库共收录数据3 143 365条，覆盖了UniProt数据库中2 278 333条不同的蛋白质序列。SWISS-MODEL数据库允许用户对数据库中的模型进行质量评价，允许用户搜索另外一种可变模板结构（alternative template structures），用户还可以使用SWISS-MODEL工作平台（http://swissmodel.expasy.org/workspace/）构建蛋白质的三维模型。最后对结构模型的注释信息，包括功能信息，可通过与其他数据库进行交叉链接得到，通过这些链接，用户就可以在蛋白质序列数据库和结构数据库之间自由切换。

2. 中国蛋白质结构数据库　中国蛋白质结构数据库存储了中国人提交的蛋白质的PDB数据（http: //lifecenter.sgst.cn/cnpdb/cn/pdbHome.do）。截至2009年7月，该数据库中总记录数58706条。点击主页（图4-8）中左侧对应链接可进行数据浏览及下载。

图4-8　中国蛋白质结构数据库主页

该数据库可根据蛋白质的ID对数据进行浏览,包括文献名、作者名及相应PubMed ID等内容,还可以用Jmol查看该蛋白质的结构。FTP站点上提供了按中国作者所占百分比整理的PDB数据的打包下载。

随着测序技术和预测方法不断发展,涌现了很多蛋白质结构相关的数据库。这些数据库存储蛋白质序列、分类、家族、二级或三级结构、膜蛋白、结构域以及结构修饰等信息(表4-6)。

表4-6 常用蛋白结构数据库

数据库	说明	网址链接
PDB	蛋白质三维结构	http: //www.rcsb.org/pdb
REAID	蛋白质结构修饰数据库	http: //pir.georgetown.edu/cgi-bin/resid
中国蛋白质结构数据库	中国蛋白质结构数据库	http: //lifecenter.sgst.cn/cnpdb/cn/pdbHome.do
BMRB	生物磁共振数据库	http: //www.bmrb.wisc.edu/
SWISS-3DIMAGE	三维结构图示	http: //us.expasy.org/sw3d/
DSSP	蛋白质二级结构参数	http: //www.cmbi.kun.nl/gv/dssp/
SWISS-MODEL	从序列模建结构	http//www.expasy.org/swissmod/SWISS-MODEL.html
FSSP	已知空间结构的蛋白质家族	http: //www.bioinfo.biocenter.helsinki.fi
SCOP	蛋白质分类数据库	http: //scop.mrc-lmb.cam.ac.uk/scop/
CATH	蛋白质分类数据库	http: //www.biochem.ucl.ac.uk/bsm/cath/
Pfam	蛋白质家族和结构域	http: //pfam.wustl.edu/
tmbase	跨膜蛋白数据库	ftp: //ulrec3.unil.ch(/pub/tmbase)
TrEMBL	EMBL的翻译数据库	http: //kr.expasy.org/sprot/
PROSITE	蛋白质功能位点	http: //kr.expasy.org/prosite/

第三节

蛋白质高级结构预测方法

Section 3　Method of Protein Advanced Structure Prediction

蛋白质结构的直接获取仍然存在瓶颈，大量序列已知蛋白质的三维结构尚未被实验方法测定出来。在这种情况下，充分利用一级序列信息和已知蛋白质的空间结构信息来预测未知蛋白质的空间结构，已经成为研究和理解蛋白质结构-功能关系的最重要手段之一。人们要求对蛋白质结构分类不仅应能够自动化处理，而且应同时具有更加准确和更低计算量等特点。蛋白质结构自动分类问题可以被纳入模式识别的范畴，通过提取分析蛋白质结构的关键特征，挖掘蕴含于大量已知类别和结构蛋白质中的结构和功能知识来构造分类器，最终实现对未知蛋白质结构的分类预测。蛋白质折叠分类识别的特征提取对象，也逐渐从序列向结构过渡。根据特征的来源，当前的研究方法可分为三类：基于序列、基于结构，以及两者混合的特征提取方法。为了从蛋白质序列或结构中获取包含更多的结构或功能信息的特征，人们通常从多个方面去提取特征，然后将所得到的各种特征组合在一起进行分类。

一、蛋白质二级结构预测方法及软件

蛋白质二级结构的预测通常被认为是蛋白结构预测的第一步，是根据它们被预测的局部结构，对蛋白序列中的氨基酸进行分类。二级结构的预测方法通常分为多序列列线预测和单序列预测的方法。由于单序列预测所提供的信息只是残基的顺序而没有其空间分布的信息，所以单序列预测的算法预测准确率并不高。多序列列线预测和神经网络的应用大大提高了二级结构预测的准确度，通过对序列比对的预测可以明确的提供单一位点在三维结构上的信息。通常二级结构预测的准确率比单序列预测能够提高10%，很多方法甚至可达到70%~77%的准确度。

（一）二级结构预测方法

1. 经验参数法　经验参数法是Chou和Fasman提出的，是一种基于单个氨基酸残基统计的经验预测方法。通过统计分析，获得每个残基出现于特定二级结构构象的形象性因子，进而利用这些倾向性因子预测蛋白质的二级结构。它使用氨基酸物理化学数据中派生出来的规律来预测二级结构。首先统计出20种氨基酸在α螺旋、β折叠和无规则卷曲中出现频率的大小，然后计算出每一种氨基酸在这几种构象中的构象参数Px，构象参数值的大小反映了该种残基出现在某种构象中的倾向性的大小。Chou和Fasman根据残基的倾向性因子提出二

级结构预测的经验规则,根据蛋白序列寻找二级结构的成核位点和终止位点。这种方法可能能够正确反映蛋白质二级结构的形成过程,但预测成功率并不高,仅有50%左右。

2. GOR算法　GOR算法是一种单序列预测方法,因其作者Garnier、Osguthorpe和Robson而得名。基于信息论和贝叶斯统计学方法,将蛋白质序列作为一连串的信息值处理。该方法不仅考虑被预测位置本身氨基酸残基的种类对该位置构象的影响,也考虑相邻残基种类对该位置构象的影响。GOR方法的具体做法是: 将序列中的每一个残基与和它的N端紧邻的8个残基以及和它 C端紧邻的8个残基一起考虑,通过对已知二级结构的蛋白样本的分析,计算出中心残基的二级结构分别为螺旋、折叠和转角时每种氨基酸出现在窗口中各个位置的频率,产生一个17×20的得分矩阵。然后预测序列中每个残基形成这些二级结构的概率。这样使预测的成功率提高到65%左右。

3. 多序列列线预测　对序列进行多序列比对,并利用多序列比对的信息进行结构的预测。调查者可找到和未知序列相似的序列家族,然后假设序列家族中的同源区有同样的二级结构,预测不是基于一个序列而是一组序列中的所有序列的一致序列。

4. 神经网络方法　神经网络算法通常是由三层相同的神经元构成的层状网络,使用反馈式学习规则,底层为输入层,中间为隐含层,顶层是输出层,信号在相邻各层间逐层传递,不相邻的各层间无联系,在学习过程中根据输入的一级结构和二级结构的关系的信息不断调整各单元之间的权重,最终目标是找到一种好的输入与输出的映像,并对未知二级结构的蛋白进行预测。神经网络方法的优点是应用方便,获得结果较快较好; 主要缺点是没有反映蛋白的物理和化学特性,而且利用大量的可调参数,使结果不易理解。许多预测程序如PHD、PSIPRED等均结合利用了神经网络的计算方法。

5. 基于已有知识的预测方法　预测方法包括Lim和Cohen两种方法。Lim方法是一种物理化学的方法,它根据氨基酸残基的物理化学性质,包括: 疏水性、亲水性、带电性以及体积大小等,并考虑残基之间的相互作用而制订出一套预测规则。对于小于50个氨基酸残基的肽链, Lim方法的预测准确率可以达到73%,另一种是 Cohen方法,它的提出当时是为了α/β蛋白的预测,基本原理是: 疏水性残基决定了二级结构的相对位置,螺旋亚单元或扩展单元是结构域的核心, α 螺旋和 β 折叠组成了结构域。

6. 混合方法　将以上几种方法选择性的混合使用,并调整它们之间使用的权重可以提高预测的准确率,目前预测准确率在70%以上的都是混合方法,其中,同源性比较方法、神经网络方法和 GOR方法应用最为广泛。

(二) 蛋白质结构域识别方法

蛋白质结构域是具有特定功能的基本结构单元。它既是蛋白质结构化分类的基础,又与蛋白质进化密切相关。它对于人们认识蛋白质的结构、功能和进化有着重要的意义。因此,蛋白质结构域的研究已成为生物信息学中的一个重要问题。通过专家手工来确定蛋白质结构域是非常可靠的。然而处在数据量急速增长的后基因组时代,人类专家的处理能力已无法满足数据分析的需要,这时自动化的预测方法则显得尤为重要。自动化的结构域预测方法可分为基于模板的方法和从头预测的方法。尽管基于模板的方法已经取得了较大的成功,但它在缺乏相应的模板信息时就不再有效。仅从序列信息来预测结构域的方法(从头预测)成为结构生物学和序列分析中的一个重要的问题。目前许多机器学习方法,如隐马尔可夫

模型、神经网络、支持向量机等已经被应用于蛋白质结构域边界的从头预测中。

1. 递归的神经网络　可使用的模型有基于长短记忆(long short-term memory，LSTM)递归网络的蛋白质结构域边界预测模型——IPSP-LSTM。该模型通过选择性记忆的递归方法对蛋白质序列中的长程相关性进行建模。该模型在整体结构域预测和多域蛋白质链的预测中的效果较好。在双域的预测中的敏感性和特异性更加平衡。

2. 支持向量机　支持向量机的基本原理是：首先通过将种子序列与数据库中已知的序列相比较，生成多序列比对结果，对比对结果进行特征提取，这些特征能够直接或间接的反映蛋白质的结构属性及结构域信息，再运用信息论的方法将特征值信息最大化。使用支持向量机学习系统对提取的特征值进行分类，实现了从多变量到单分类结果的非线性映射。

(三) 二级结构预测相关软件

目前较为常用的二级结构预测软件PSIPRED、Jpred、PREDATOR、PSA和SOMPA等都有在线服务器；进入这些软件的主页，输入Fasta格式的目的蛋白序列，在网页上直接选取适合的蛋白质结构预测算法，点submit运行即可。

1. Jpred　Jpred 是一种蛋白质二级结构预测网络服务器，由 Barton Group创建于1998年。通过提交单一蛋白质序列或多重蛋白质序列并运行，Jpred就可以预测出蛋白质序列的二级结构：α-螺旋、β-折叠或无规则卷曲。Jpred应用了Jnet 神经网络算法，准确率达到了76.4%。其基本原理是：Jpred服务器通过DSC、PHD、NNSSP、PREDATOR、ZPRED和MULPRED六种预测方法进行预测，它们都采用了多重序列的进化信息。NNSSP依据最大同源性，PDH采用神经网络，DSC根据线性识别，MULPRED 联合不同的单一序列预测方法，PREDATOR 考虑氢键倾向性，ZPRED 加权预测。最后将六个结果总结为一个简单的文件格式。

2. SOPMA　位于法国里昂的CNRS(centre national de la recherche scientifique)(http://pbil.ibcp.fr/)使用独特的方法进行蛋白质二级结构预测。它是使用5种相互独立的方法进行预测，并将结果汇集整理成一个"一致预测结果"。这5种方法包括：Garnier-Gibrat-Robson(GOR)方法、Levin同源预测方法、双重预测方法、PHD方法和CNRS自己的SOPMA方法。简单地说，SOPMA这种自优化的预测方法建立了已知二级结构序列的次级数据库，库中的每个蛋白质都经过基于相似性的二级结构预测。然后用次级库中得到的信息去对查询序列进行二级结构预测。

3. nnPredict　nnpredict(http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html)算法使用了一个双层、前馈神经网络去给每个氨基酸分配预测的类型。在预测时，服务器使用FASTA格式的文件，其中有单字符或三字符的序列以及蛋白质的折叠类(α、β或α/β)。残基被分为几类，如α螺旋(H)、β链(E)或其他(-)。若对给定残基未给出预测，则会标上问号(?)，这说明无法做出可信的分配。若没有关于折叠类的信息，预测也能在不定折叠类的情况下进行，而且这是缺省的工作方式。据报道，对于最佳实例的预测，nnpredict的准确率超过了65%。

4. PredictProtein　PredictProtein(http://cubic.bioc.columbia.edu/predictprotein/)在预测中应用了略为不同的方法。首先，蛋白质序列被作为查询序列在SWISS-PROT库中搜索相似的序列。当相似的序列被找到后，一个名为MaxHom的算法被用来进行一次基于特征简图的

多序列比对。MaxHom用迭代的方法来构造比对：当第一次搜索SWISS-PROT后，所有找到的序列与查询序列进行比对，并构造出一个比对后的特征简图。然后，这个简图又被用来在SWISS-PROT中搜索新的相似序列。由MaxHom产生的多序列比对随后被置入一个神经网络，用PHD的方法进行预测。

【例4-2】在SOPMA中预测人类驱动蛋白的二级结构

（1）进入SOPMA主页：

（http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl ? page=/NPSA/npsa_sopma.html）；

（2）如图4-9所示，在“Paste a protein sequence below”下的空白处提交人类驱动蛋白序列，设置拟定的参数，点击“SUBMIT”按钮进行分析；

SOPMA SECONDARY STRUCTURE PREDICTION METHOD

[Abstract] [NPS@ help] [Original server]

Sequence name (optional) :

Paste a protein sequence below : help

Output width : 70

SUBMIT　CLEAR

Parameters

Number of conformational states : 4 (Helix, Sheet, Turn, Coil)

Similarity threshold : 8

Window width : 17

图4-9　SOPMA首页

（3）结果如图4-10；SOPMA方法预测的二级结构主要含有α-螺旋（h），占37.44%；延伸链（e），占14.26%；β-折叠（t），占4.16%；无轨卷曲（c），占44.13%。

5. 蛋白质二级结构其他预测软件　目前，还有很多蛋白质二级结构在线预测软件，如APSSP、CFSSP、PROF和PSIPRED等（表4-7）。并非所有的方法都是默认执行的，有些方法，如跨膜螺旋的预测，在自动运行时使用特殊的保守起始值，而在有明确要求时使用不同的起始值。以下方法可选择使用：MaxHom、BLASTP、PSI-BLAST、SEG、PHDsec、PHDacc、PHDhtm，PROFsec、PROFacc、COILS、CYSPRED、ASP、PROSITE、ProDom、CHOP、NORSp、PROFtmb、PROFcon08、LOCkey、LOChom、PredictNLS 和 LOCnet。使用者可以明确要求使用TOPITS+或用EvalSec评估二级结构预测方法的准确率。注意，某些方法的使用有以下优势：加快执行的速度和简化结果。然而请记住，数据库搜索及其结果是速度和结果字节数的限制因素。

SOPMA result for : kinesin

Abstract Geourjon, C. & Deléage, G., SOPMA: Significant improvement in protein secondary structure prediction by consensus prediction from multiple alignments., Cabios (1995) 11, 681-684

View SOPMA in: [AnTheProt (PC) , Download...] [HELP]

```
        10        20        30        40        50        60        70
         |         |         |         |         |         |         |
MASQFCLPESPCLSPLKPLKPHFGDIQEGIYAAIQRSDKRIHLAVVTEINRENYWVTVEWVEKAVKKGKK
hhhhecccccccccccccccccchhhhtteeeeeeccttcchheeeehhcttcceeeeeehhttccttcc
IDLETILLLNPALDSAEHPMPPLPLSPLALAPSSAIRDQRTVTKWVAMIPQKNQTASGDSLDVRVPSKPC
eehhheeeccttcccccccccccccccccccccccccccccccccccccccccccccccccccccccccc
LMKQKKSPCLWEIQKLQEQREKRRRLQQEIRARRALDVNTRNPNYEIMHMIEEYRRHLDSSKISVLEPPQ
ccccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhccccccchhhhhhhhhhhhccccccccccccccc
EHRICVCVRKRPLNQRETTLKDLDIITVPSDNVVMVHESKQKVDLTRYLQNQTFCFDHAFDDKASNELVY
ccceeeeeeecccccccccccceeeeecttceeeeeeccccchhhhhhhhttceeeeeeeccccchhhh
QFTAQPLVESIFRKGMATCFAYGQTRSGKTYTVGGDFSGTAQDCSKGIYALVAQDVFLLLRNSTYEKLDL
hhhcchhhhhhhttcceeeeeeccccccceeeeeccccccccccctttcehhhhhhhhhhhhcccccccch
KVYGTFFEIYGGKVYDLLNWKKKLQVLEDGNQQIQVVGLQEKEVCCVEEVLNLVEIGNSCRTSRQTSVNA
eeeeehhhhhhhhhhhhhhttcceeeehttccceeeeecchhhhcchhhhhhhhhttccccccccccccc
HSSRSHAVFQIILKSGGIMHGKFSLVDLAGNERGADTTKASRKRQLEGAEINKSLLALKECILALGQNKP
ccccchheeeeeehccccccceeeeeeecccccccccccccccccccchhhhhhhhhhhhhhhhhhhcccc
HTPFRASKLALVLRDSFIGQNSSTCMIATISPGMTSCENTLNTLRYANRVKKLNVDVRPYHRGHYPIGHE
ccccccchhhhhhhhhcccccccceeeeeeccttccchhhhhhhhhhhhhhhhheeccccccccccccccc
APRMLKSHIGNSEMSLQRDEFIKIPYVQSEEQKEIEEVETLPTLLGKDTTISGKGSSQWLENIQERAGGV
chhhhhhhhhcccccccccccceecccccchhhhhhhhhhhhhhhhhhhhhhhhhccchhhhhhhhhhhhhhc
HHDIDFCIARSLSILEQKIDALTEIQKKLKLLLADLHVKSKVE
ccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhccee
```

Sequence length :   673

SOPMA :

Alpha helix	(Hh) :	252	is	37.44%
3_{10} helix	(Gg) :	0	is	0.00%
Pi helix	(Ii) :	0	is	0.00%
Beta bridge	(Bb) :	0	is	0.00%
Extended strand	(Ee) :	96	is	14.26%
Beta turn	(Tt) :	28	is	4.16%
Bend region	(Ss) :	0	is	0.00%
Random coil	(Cc) :	297	is	44.13%
Ambigous states	(?) :	0	is	0.00%
Other states	:	0	is	0.00%

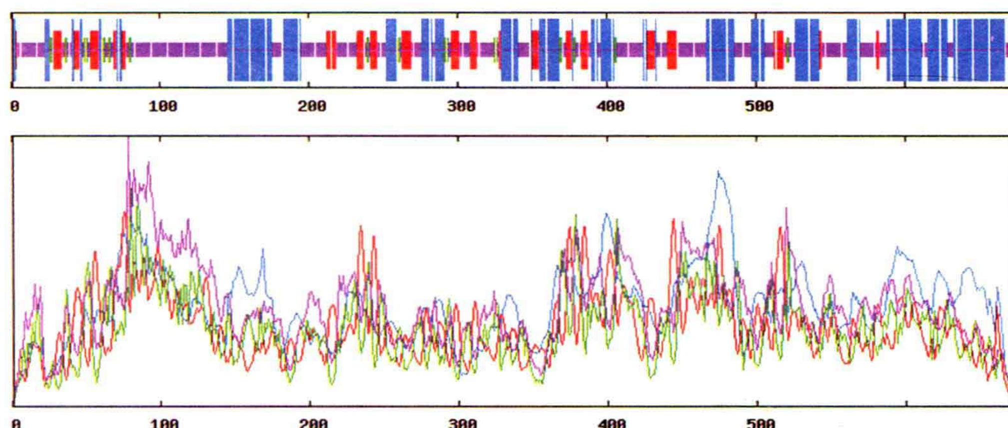


图4-10　SOPMA预测结果

表4-7　常用二级结构预测软件

软件	网址	方法
APSSP	http: //imtech.res.in/raghava/apssp/	基于最近邻和神经网络方法，根据氨基酸序列预测蛋白质的二级结构
CFSSP	http: //www.biogem.org/tool/chou-fasman/	Chou & Fasman算法，根据氨基酸序列预测蛋白质二级结构
PROF	http: //www.aber.ac.uk/~phiwww/prof/	根据氨基酸多重序列比对预测蛋白质二级结构
PSIPRED	http: //bioinf.cs.ucl.ac.uk/psipred/	提供跨膜蛋白拓扑结构预测和蛋白profile折叠结构识别工具

二、蛋白质三级结构的预测方法及软件

(一) 蛋白质三级结构的预测方法

目前,蛋白质三级结构的预测方法主要有三类:同源模建法、折叠识别法和从头预测。

1. 同源模建法　同源模建法也称为比较模建法(comparative modeling)。同源模建法的基础是同源蛋白质空间结构比蛋白质序列保守性更强的理论,基本假设是蛋白质结构具有某种规则性,其可能的空间结构的基本形态种类有限,各个形态由各物种特定的氨基酸序列所决定。在蛋白质序列的一致性大于30%的前提下,一个未知结构的蛋白质可以利用一个或一个以上与其相关的蛋白质结构来建立其空间结构。一般来说,目标蛋白质序列和模板序列的相似性越高,所模建出来的结构正确性、可信度也就越高。

2. 折叠识别法　有许多蛋白质氨基酸序列大不相同,但是却拥有极为相似的三维结构,在这种情况下同源模建法因为序列一致性太低而失效,因此,一些科学家还提出了一种预测蛋白质三级结构的新策略,这类方法被称为Threading 方法或折叠类型识别方法,这一方法的基本思想是假定被预测蛋白质的折叠类型与某一已知结构的蛋白质的折叠类型相同,这样,蛋白质结构预测的问题就转变为与已知空间结构的蛋白质比对,从而大大减少了预测蛋白质结构的难度,而且不需要预测二级结构,即直接预测三级结构,从而可以避免二级结构预测不准确的限制,是一种有潜力的预测方法。

折叠识别法的实现过程是总结出已知的独立蛋白质结构模式作为可与未知结构进行匹配的模板,然后通过学习现有的数据库总结出评价序列与结构匹配优劣的平均势函数作为判别标准,选择出未知序列与已知特定结构的最佳匹配。给定一个结构未知的查寻序列及一些蛋白质的结构(或结构的片段),计算这个序列与其中某个结构的折叠匹配关系,然后将氨基酸序列和三维结构在空间中的位置做排列,再运用适当的计分方式,计算匹配得分,根据得分的高低,对序列与折叠的立体结构进行评估。

3. 从头预测　在既没有已知结构的同源蛋白质、也没有已知结构的远程同源蛋白质的情况下,上述两种蛋白质结构预测的方法都不能用,这时只能采用从头预测方法(Abinitio),即直接根据序列本身来预测其结构。从头预测方法一般由下列3个部分组成:①一种蛋白质的几何表示方法:由于表示和处理所有原子和溶剂环境的计算开销非常大,因此需要对蛋白质和溶剂的表示形式作近似处理,例如,使用一个或少数几个原子代表一个氨基酸残基。②一种能量函数及其参数,或者一个合理的构象得分函数,以便计算各种构象的能量。通过对已知结构的蛋白质进行统计分析,可以确定蛋白质构象能量函数中的各个参数或者得分函数。③一种构象空间搜索技术:必须选择一个优化方法,以便对构象空间进行快速搜索,迅速找到与某一全局最小能量相对应的构象。其中,构象空间搜索和能量函数的建立是从头预测方法的关键。

(二) 蛋白质三级结构预测的软件

1. SWISS-MODEL(http://www.expasy.ch/swissmod/SWISS-MODEL.html)是自动的蛋白质同源模建服务器。程序先把提交的序列在ExPdb晶体图像数据库中搜索相似性足够高的

同源序列，建立最初的原子模型，再对这个模型进行优化产生预测的结构模型。

由于比较建模程序可以具有不同的复杂性，该服务主要有以下三种方式：

（1）简捷模式（First Approach mode）：这种模式提供一个简捷的用户界面：用户只需要输入一条氨基酸序列，服务器就会自动选择合适的模板。或者，用户也可以自己指定模板（最多5条），这些模板可以来自ExPDB模板数据库，也可以是用户选择的含坐标参数的模板文件。如果一条模板与提交的目标序列相似度大于25%，建模程序就会自动开始运行。但是，模板的可靠性会随着模板与目标序列之间相似度的降低而降低，如果相似度不到50%往往就需要用手工来调整序列比对。这种模式只能进行大于25个残基的单链蛋白三维结构预测。

（2）比对界面（Alignment Interface）：这种模式要求用户提供两条已经比对好的序列，并指定哪一条是目标序列，哪一条是模板序列（模板序列应该对应于ExPDB模板数据库中一条已经知道其空间结构的蛋白序列）。服务器会依据用户提供的信息进行建模预测。

（3）工程模式（Project mode）：手工操作建模过程：该模式需要用户首先构建一个DeepView工程文件，这个工程文件包括模板的结构信息和目标序列与模板序列间的比对信息。这种模式可以让用户控制许多参数，例如，模板的选择，比对中的缺口位置等。此外，这个模式也可以用于"first approach mode简捷模式"输出结果的进一步加工完善。

此外，SWISS-MODEL还具有其他两种内容上的模式：①Oligomer modeling（寡聚蛋白建模）：对于具有四级结构的目标蛋白，SWISS-MODEL提供多聚模板的模式，用于多单体的蛋白质建模。这一模式弥补了简捷模式中只能提交单个目标序列，不能同时预测两条及以上目标序列的蛋白三维结构的不足；②GPCR mode（G蛋白偶联受体模式）：是专门对7次跨膜G蛋白偶联受体的结构预测。

【例4-3】用SWISS-MODEL自动方式以人类驱动蛋白序列为例说明三级结构建模过程

第一步：进入SWISS-MODEL三级结构预测服务器主页（图4-11）；.

SIB Swiss Institute of Bioinformatics | BIOZENTRUM Universität Basel | SWISS-MODEL

Modelling
myWorkspace
Automated Mode
Alignment Mode
Project Mode

Tools
Template Identification
Domain Annotation
Structure Assessment
Template Library

Repository
Search by Sequence
Search by AC
Search by full text

SWISS-MODEL is a fully automated protein structure homology-modeling server, accessible via the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer). The purpose of this server is to make Protein Modelling accessible to all biochemists and molecular biologists worldwide.

What's new?
- New Beta SWISS-MODEL pipeline for automated model building with oligomers and ligands is now available for testing.
- Find more news on SWISS-MODEL Blog
- ... faster news on Twitter
- Follow us on Facebook

SWISS-MODEL Team
Torsten Schwede: Project Leader
Florian Kiefer: SWISS-MODEL Repository
Lorenza Bordoli: Method Development and user support
Konstantin Arnold: SWISS-MODEL Workspace

References:
When you publish or report results using SWISS-MODEL, please cite the relevant publications:
- Arnold K., Bordoli L., Kopp J., and Schwede T. (2006). The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling. Bioinformatics, 22,195-201.
- Kiefer F, Arnold K, Künzli M, Bordoli L, Schwede T (2009). The SWISS-MODEL Repository and associated resources. Nucleic Acids Research. 37, D387-D392.
- Peitsch, M. C. (1995) Protein modeling by E-mail Bio/Technology 13: 658-660.

图4-11　SWISS_MODEL预测服务器主页

第二步：选择“Automated Mode”→粘入从NCBI上搜索到的人类驱动蛋白kinesin蛋白质序列；在这里可以填写E-mail地址，将结果发送至电子邮箱，也可以在新的网页上直接展示结果；

第三步：点击“Submit Modeling Request”即可；

第四步：直接在页面上查看蛋白质kinesin的三级结构信息（图4-12）；

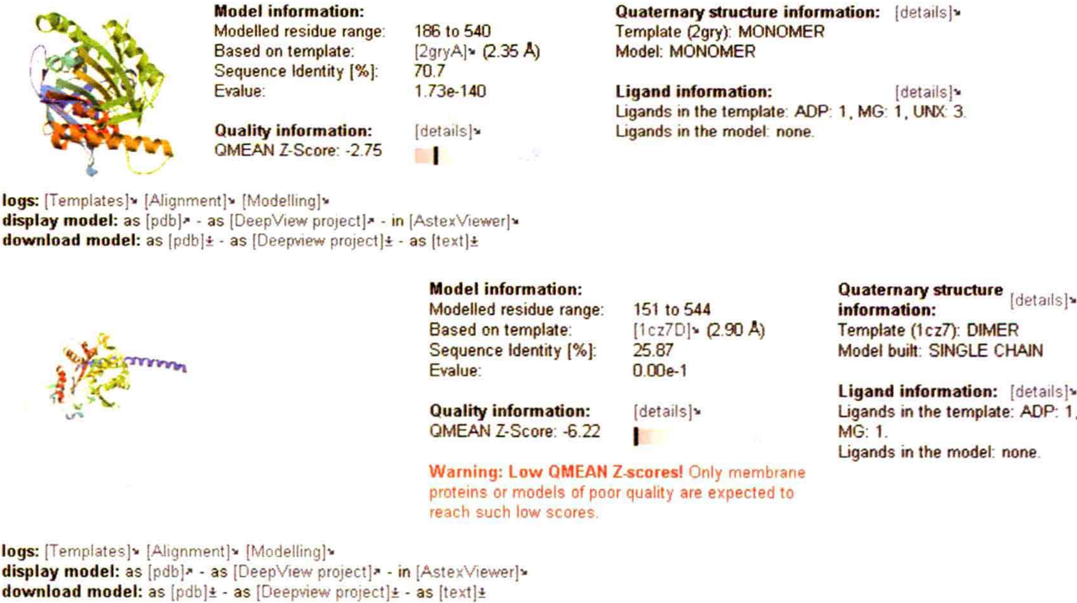


图4-12　借助SWISSA-MODEL查找与kinesin三级结构相似模型

第五步：结果分析发现系统自动选用的是其中相似性最高的两个模型，分别是2gry和1cz7，从图4-12中可看出其三级结构含有α-螺旋和平行β-折叠链。从模板信息里可以得到所模拟目标蛋白的残基范围、所用模板、序列相似性及E值。另外，可通过展示模型获得模板的具体信息或者通过下载模板保存其三级结构的PDB格式。

第六步：使用PHYRE 工具（http: //www.sbg.bio.ic.ac.uk/~phyre/index.cgi）查看蛋白质二级结构的比对细节、同源性结构等信息（图4-13）；

第七步：使用CBS（http: //www.cbs.dtu.dk/index.shtml）中的FeatureMap3D直接对蛋白质序列做基于PDB数据库的蛋白质三维结构图（图4-14）；

第八步：预测结果可以使用显示生物大分子三维结构图像的软件，如RasMol、PyMol、Cn3D、SWISS-pdbVeiwer等显示（图4-15）。

2. PROCARB（http: //www.procarb.org/）是一款可预测糖蛋白的软件，其包含的同源建模模块是基于同源建模的方法预测糖蛋白的三维结构。预测的过程是：首先，由于糖蛋白分为N连接糖蛋白和O连接糖蛋白，因此，分别在Swissprot数据库中查找N连接糖蛋白，在O-glycbase数据库中查找O连接糖蛋白。其次，在搜索到的糖蛋白中，选择与蛋白家族的序列相似性在30%以上的家族中的一个蛋白作为模型。在被选择的糖蛋白序列中，要求至少有一个糖基化位点是在Swissprot中存在和注释的。然后，将序列输入到3D-JIGSAW（http: //bmm.cancerresearchuk.org/~3djigsaw/）服务器中对蛋白质进行建模。最后，使用CHARMm（http: //

Secondary Structure Prediction

Index	. . . . 10. . . . . 20. . . . . 30. . . . . 40. . . . . 50. . . . . 60. . . . . 70. . . .
Query Sequence	MASQFCLPESPCLSPLKPLKPHFSDIQECIYAAIQRSDKRIHLAVVTEINRENYWVTVEWVEKAVKKGKKIDLE
psipred	eeeeeee eeeeeeeeeee eeeeee e ee hh
jnet	hhhhhhhhh hh hhhhh eeeeeeeeeeee eeeeeeeee ee eeeee ee hhh
sspro	eeeeeeeeeeeee eeeeeeeee e eeeee e e ee hh
Consensus	eeeeeeeeeee eeeeeeeee e eeeee e hh
Cons_prob	865445666656777666667777556645899996599588899998758985899997668855776455 7

Disorder Prediction

Index	. . . . 10. . . . . 20. . . . . 30. . . . . 40. . . . . 50. . . . . 60. . . . . 70. . . .
Disopred	ddddddd ooo oo dd ooooo d oooooooooo oooooooooo oooooooooo ddd oo ooooo oooooooooo oooo
Diso_prob	99988774324456214335432243332413211121214432244444765410011000221311130013
Prosite	

FINALLY - a new fold library has been installed! Let me know if you have any problems.

To predict functional residues and GO classification, try ConFunc

Fold Recognition

View Alignments	SCOP Code	View Model	E-value	Estimated Precision	BioText	Fold/PDB descriptor	Superfamily	Family	(beta-test)
	c1v8kA_ (length:410) 60% i.d.	Jmol MDL	9e-32	100 %	n/a	**PDB header**:structural protein	**Chain**: A: **PDB Molecule**:kinesin-like protein kif2c;	**PDBTitle**: the crystal structure of the minimal functional domain of the2 microtubule destabilizer kif2c complexed	n/a

图4-13 蛋白质二级结构的比对细节、同源性结构

FeatureMap3D - query result

Please cite:

FeatureMap3D - a tool to map protein features and sequence conservation onto homologous structures in the PDB
Rasmus Wernersson, Kristoffer Rapacki, Hans-Henrik Stærfeldt, Peter Wad Sackett, and Anne Mølgaard
Nucl. Acids Res. 2006 34: W84-W88

A guide of how to read the result can be found here: output.

Result for Seq1.1.1 (673 aa)

Download: GetStruct report | PDB structure (2GRY) | 3D plot (png) | PyMol script | All files (TAR archive) | All files (ZIP archive)



图4-14 预测的kinesin三维模型与模板的比对

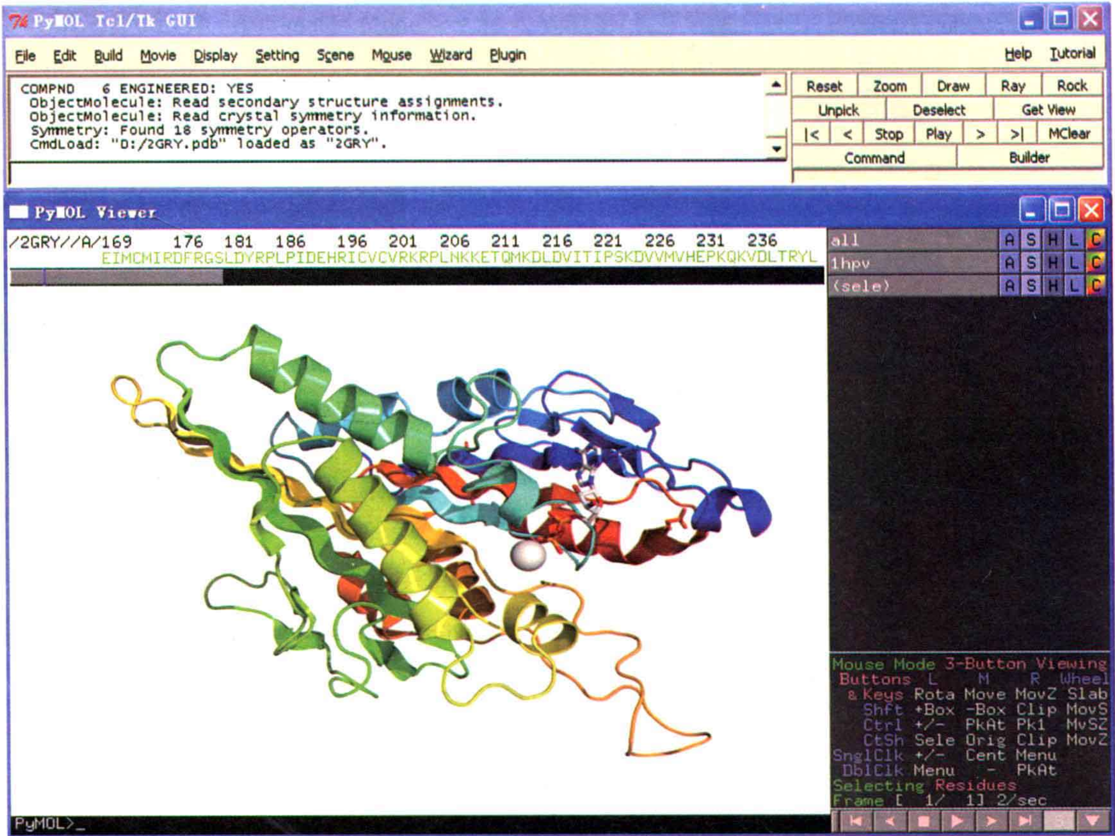


图4-15 kinesin三维结构预测结果可视化

www.charmm.org/)对模型进行优化。

【例4-4】应用PROCARB预测人类酸性[神经]酰胺酶的三维结构

(1)打开PROCARB(http: //www.procarb.org/),进入主页(图4-16);

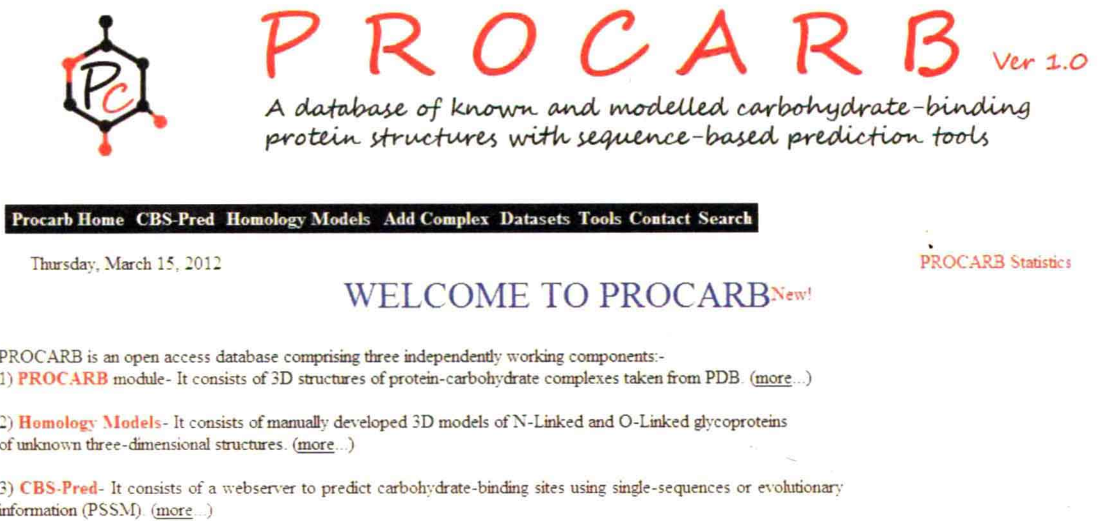


图4-16 PROCARB的首页

（2）点击“HUMOLOGY MODELS”进入同源建模模块（图4-17），在“Enter Swissprot ID”输入框中输入人类酸性[神经]酰胺酶的ID“Q13510”，点击“Enter”，可自动连接3D-JIGSAW同源建模服务器，进行三维结构的预测；

HOMOLOGY MODELS

Here, you can find the three dimensional structure models of diverse types of glycoproteins.
All these models were automatically generated by using 3DJIGSAW homology modeling server.

Enter Swissprot ID : Q13510 | Enter | Reset | OR Keyword Search

图4-17　PROCARB的UMOLOGY MODELS模块

（3）获得人类酸性[神经]酰胺酶的基本信息和模建的三维结构信息等（图4-18）。例如，蛋白质功能、糖基化位点、Pfam的描述和模建的3D结构的下载链接等。

Uniprot ID	Q13510
Protein Name	Acid ceramidase
Source	Homo sapiens
Pfam Description	Linear amide C-N hydrolases, choloylglycine hydrolase family
Gene Name	ASAH1
Function	It hydrolyzes the sphingolipid ceramide into sphingosine and free fatty acid.
Glycosylation Sites	ASN259 & ASN286
Model 3D Structure	Download
PubMed Central	Click Here to search articles in PMC for Acid ceramidase

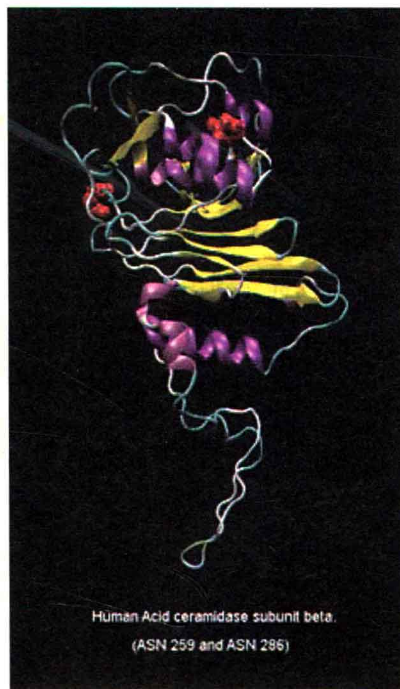


图4-18　人类酸性[神经]酰胺酶的预测结果

3. Phyre　Phyre（http: //www.sbg.bio.ic.ac.uk/phyre/html/index.html）主要利用折叠方式对模型进行预测，主要是针对网上数据库中没有高同源性模板的蛋白三级结构的预测。Phyre服务器是2005年发布的，其原理是基于每个蛋白特异的位点打分矩阵进行profile-profile比对。Phyre2服务器是2011年发布的，它增加了一些功能：如升级的界面，全面更新的折叠库和使用HHpred / HHsearch软件包预测同源性等。

4. 其他蛋白质三维结构预测软件　目前，还有很多蛋白质三级结构常用预测软件，如CPHmodels、ESyPred3D、LOOPP、FUGUE和HMMSTR等（表4-8）。

表4-8　三级结构其他预测软件

软件	网址	简介
CPHmodels	http: //www.cbs.dtu.dk/services/CPHmodels/	基于profile–profile的比对的功能打分算法和远源的同源蛋白模型算法方法预测蛋白质三级结构
ESyPred3D	http: //www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/	基于神经网络算法提高同源建模准确性的预测
LOOPP	http: //cbsuapps.tc.cornell.edu/loopp.aspx	基于多个信号整合成一个打分的方法的折叠识别算法
FUGUE	http: //tardis.nibio.go.jp/fugue/	扫描数据库的结构谱，计算序列结构的适应性打分，通过序列和结构比对预测蛋白质折叠
HMMSTR	http: //www.bioinfo.rpi.edu/~bystrc/hmmstr/server.php	基于隐马尔科夫模型预测蛋白质三级结构

三、对结构预测结果的评价

面对多种的模型和预测方法，有多种公共范围的实验评估方法，主要是LB、CASP和CAFASP、EVA等方法。

1. EVA　哥伦比亚大学的研究者们提供了一种以连续的、自动化、大规模的工作方式进行蛋白质结构预测算法评估的Web服务器EVA（http: //cubic.bioc.columbia.edu/eva）。目前，EVA评估了一系列在网上可获得的预测算法的表现。每周，最新被测定结构的蛋白质的序列被自动提交到预测服务器，然后返回评测结果，并形成摘要，在网上发布。

2. CASP　CASP是在大规模实验的基础上对蛋白质结构预测进行测评的方法。测评工作分为三步：从实验研究协会收集并确定预测目标蛋白，从结构模型研究协会获得预测结果，讨论和测评。相关的具体结构由X射线衍射晶体检测学家和磁共振波谱学家提供。预测目标蛋白涉及了三个预测领域：模型比对，折叠识别和从头预测方法。

3. LiveBench（LB）实验方法　该实验方法由Rychlewski和Fischer创建。每周收集新公布的蛋白质结构，利用这些相对大量的预测靶，LB不断地对各自动服务器进行能力评估，约半年评估这些预测方法一次。

另外，对蛋白质三维结构的实验或理论模型进行检查以发现可能错误的还有其他的方法，如PROCHECK和WHAT_CHECK等。开发更复杂和自动的计算机建模方法将极大地增加结构基因组建模蛋白的范围。在该领域关键的问题包括：①对于在PDB库中相似的序列（尤其是那些与靶蛋白弱或远距离同源的）如何确定正确的模板和如何优化模板使其与天然构象相近；②若序列无合适的模板，如何从头开始进行正确拓扑的建模。

随着人们对蛋白质序列、结构、功能相互关系的更深入的了解、技术的不断进步以及新算法、新方法的呈现，基于实验和预测方法将会有越来越多的蛋白质结构被精确解析和获得。

第四节

基于蛋白质结构的功能分析

Section 4 Function Analysis Based on Protein Structure

非催化保守功能域经常介导蛋白质与蛋白质间的相互作用，这些多肽识别模块对多个蛋白质复合物的装配至关重要。功能域数据库与功能域阵列联合应用，使蛋白质间的相互作用探索更容易。蛋白质间序列相似性高于40%时，该蛋白质同其序列相似蛋白可能有某些由保守序列发挥的相同生物化学作用；蛋白质间序列保守性低于40%时，可从高级结构预测功能。蛋白质有多个功能域可对应该蛋白质的某些精细功能。从高级结构预测功能实际上是预测蛋白质的某些局部的基本生物化学作用而不是全部生物学功能。按蛋白质功能分类的数据库如SPIN-PP、MIPS等，为新蛋白功能预测提供了很多有用信息。

一、蛋白质结构与功能基础

蛋白质的空间排列在行使功能时起至关重要的作用。酶活性的研究表明，蛋白质中只有一小部分参与催化活性位点，而其余的极大部分仅用作为形成和固定活性位点的稳定基础。因此，具有不同的一级甚至三级结构的蛋白质可能具有相似乃至完全相同的生物化学功能。

在进化中保守的蛋白质高级结构通常对应某些保守的精细生物化学功能，故结构相似的蛋白质会有某些相似的精细生物化学功能。对已知结构的蛋白质进行分类，搜寻同类蛋白的功能是预测目标蛋白功能的有效手段。

最早基于结构进行蛋白质功能注释的方法是搜索与目标蛋白质结构相似的蛋白质，并将其功能转移给目标蛋白质。此过程中需要进行蛋白质的结构比对和判断结构相似程度。可将这种相似性估值转化为序列比对问题，利用序列比对经典算法来解决结构比对问题，如DaliLite、SSM、STRUCTAL、MultiProt和3DCoffee等。基于“具有相似功能的蛋白质定位于结构空间图中相邻近的位置”，Hou等（2005）使用多维度标度技术（multi-dimensional scaling，MDS）构建了一个蛋白质结构空间图（SSM），根据DaliLite结构比对方法进行相似性打分，最终在构建的结构空间中按照距离阈值将一个新的蛋白质归类到某个功能类别中。

还有一些方法试图将结构相似性方法与其他方法结合进行功能决策。例如，考虑一个系统发育上下文中的结构相似性，会增加功能注释的精确性。综合方法致力于在特定生物学背景下解决结构比对问题，有助于提高结构功能预测的精确性。

二、蛋白质结构与功能关系数据库

蛋白质结构与功能关系数据是进行蛋白质功能预测及蛋白质设计的基础。目前已有一些蛋白质结构与功能关系的数据库，如PIR、Pfam和InterPro等。

(一) Pfam数据库

Pfam(the protein families database)是通过自动比对构建的蛋白质结构域家族数据库，它收集了大量的蛋白质多重序列排布以及HMMs(profile hidden Markov models)文件的数据，将具有结构相似性的序列归为一类，可用类的名称查询到原始序列比对信息。它可广泛用于通过序列比对推测蛋白质结构域排布形式及其功能等领域。最新的Pfam 25.0版本涵盖了12 273个蛋白质家族，这些Pfam家族是基于SWISS-PROT以及TrEMBL中的蛋白质数据的。应用Wise 2软件包可以用基因组DNA对Pfam文库进行直接搜索，在地址栏中输入http: //pfam.sanger.ac.uk/ 打开Pfam(图4-19)。有多个网站支持这类数据库和搜索。



图4-19 Pfam数据库主页

Pfam数据库包含Pfam-A.seed和Pfam-A.full等文件，这些是以Stockholm格式注释的"seed"和"full"排布；PfamFrag是为搜索相匹配的蛋白片段而特别设计的HMMs文件文库；PfamB是以Stockholm格式注释的Pfam-B家族数据文件；Diff是用来对Pfam来源数据进行更新的文件；Pfamseq是以fasta格式注释的序列数据。Pfam数据库包括文本搜索、蛋白质HMM搜索、DNA HMM搜索、浏览PFAM、NIFAS和结构域查询等几个部分。可进行多种方式的搜索：①直接在JUMP TO中输入要搜索蛋白质的Pfam accesion或ID；②也可以在VIEW A SEQUENCE中输入在UniProt、NCBI或metagenomic序列数据库中已有蛋白质的序列accesion或ID；③或在SEQUENCE SEARCH中直接输入要查询的蛋白质序列，也可以搜索蛋白质的FAMILY、CLAN等。

另外，Pfam还包括蛋白质的功能注释、参考文献以及与其相应家族信息相链接的数

据库。每个Pfam家族包括“seed alignment”（由家族中具有代表性的成员构成）和“full alignment”（所有家族成员构成）两部分。所有排布都采用来源于Pfamseq的数据。在“seed alignment”基础上应用HMMER（http: //hmmer.wustl.edu）建立HMM文件对Pfamseq序列数据库进行搜索。Pfam的重要功能包括将蛋白质快速自动划分入不同的结构域家族。当前主要运用HMMer软件对蛋白质翻译进行注释，或应用Gene Wise 2软件直接预测基因并注释基因组DNA。GeneWise的检测结果表明，在同源区域内它预测基因的准确性可达到98%。结构域边界选择错误可能造成家族分类重叠或遗漏。但随着Pfam数据库的不断完善，其功能将日趋完善。

【例4-4】在Pfam数据库中查询人类原癌基因VAV

首先，进入Pfam数据库的首页，点击VIEW A SEQUENCE，输入人类原癌基因VAV编码蛋白质的UniProt ID（P15498）进行搜索（图4-20）。

wellcome trust sanger institute

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam

keyword search Go

Pfam 26.0 (November 2011, 13672 families)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. More...

QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches
VIEW A PFAM FAMILY	View Pfam family annotation and alignments
VIEW A CLAN	See groups of related families
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure
KEYWORD SEARCH	Query Pfam by keywords
JUMP TO	P15498 Go Example

Enter any type of accession or ID to jump to the page for a Pfam family or clan, UniProt sequence, PDB structure, etc.

Or view the help pages for more information

在搜索框中输入P15498

Recent Pfam blog posts

Hide this

Proposed Pfam release changes (posted 27 February 2012)

The current Pfam release, version 26.0, took approximately 4 months to nurse through the various stages of updating the sequence database, resolving overlaps between families, rebuilding the MySQL database and performing all of the post-processing that constitutes the 'release'. The production team strives to make two releases a year, but I really do not fancy [...]

The Pfam website in a virtual machine (posted 26 January 2012)

Since releasing the new Pfam website four years ago, we've had a steady trickle of mails from users who would like to install and run the site within their own local environment. It used to be possible to do just

图4-20 查询人类原癌基因VAV

查询结果如图4-21所示，左侧Summary标签页中包含VAV的基本信息：来源、长度及所包含的结构域信息。VAV共有6种7个结构域：CH、RhoGEF、PH、C1_1、SH3_1、SH2，在图中以不同颜色表示，点击各结构域的链接可以进一步查看各结构域的信息。

点击左侧Sequence标签，可显示该蛋白质的序列信息（图4-22）。

Structures标签页中显示各结构域在UniProt及PDB中的信息（图4-23），可以用三种形式查看其对应的三维结构：Jmol、AstexViewer、SPICE。

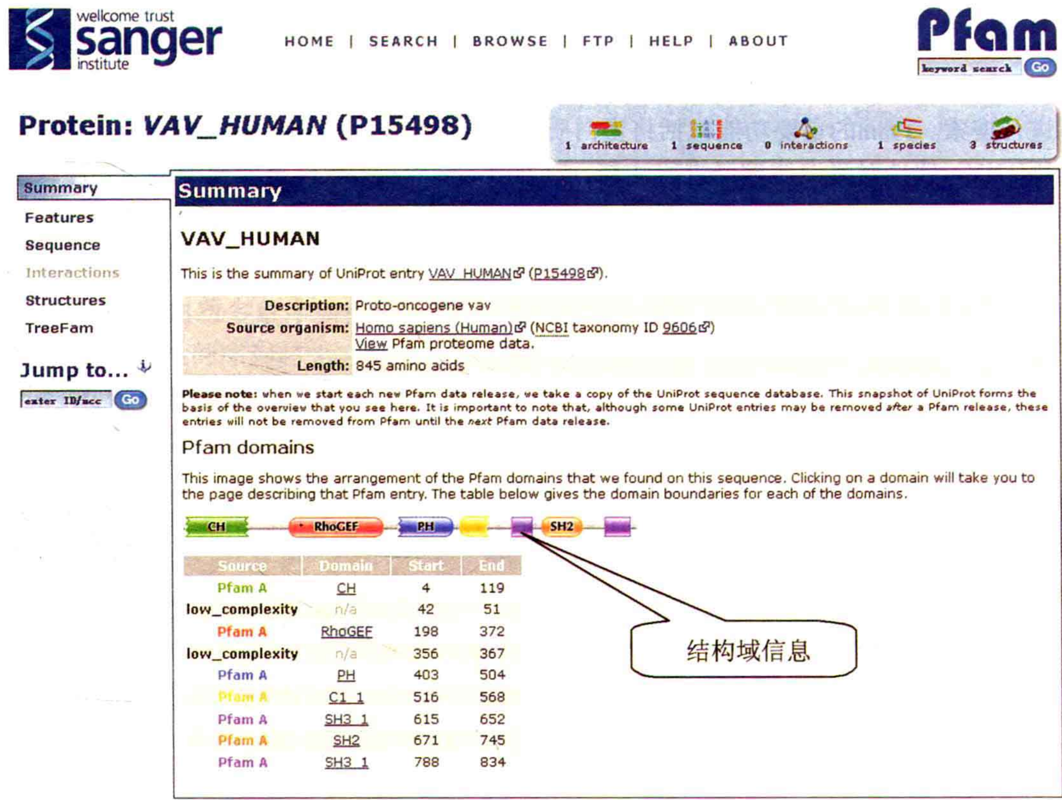


图4-21 人类原癌基因VAV的查询结果

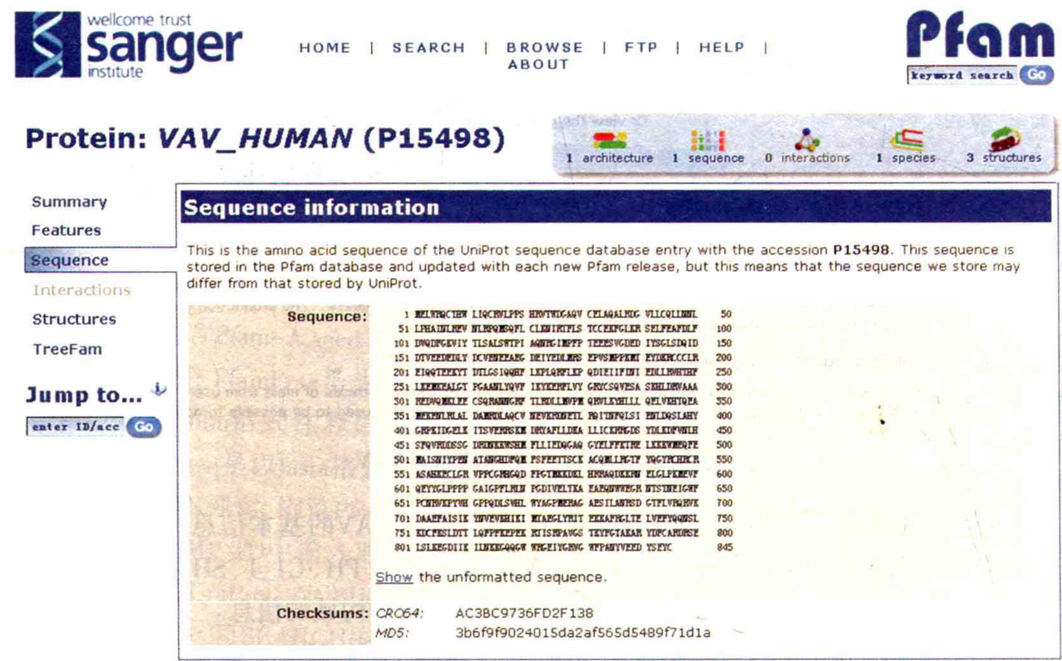


图4-22 人类原癌基因VAV的序列信息

wellcome trust sanger institute

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam

keyword search Go

Protein: *VAV_HUMAN* (P15498)

1 architecture 1 sequence 0 interactions 1 species 3 structures

Summary
Features
Sequence
Interactions
Structures
TreeFam

Jump to...

enter ID/acc Go

Structures

For those sequences which have a structure in the Protein DataBank, we use the mapping between UniProt, PDB and Pfam coordinate systems from the MSD group, to allow us to map Pfam domains onto UniProt three-dimensional structures. The table below shows the mapping between Pfam domains, this UniProt entry and a corresponding three dimensional structure.

Pfam family	UniProt residues	PDB ID	PDB chain ID	PDB residues	View
C1_1	516 - 568	3KY9	A	516 - 568	Jmol AstexViewer SPICE
			B	516 - 568	Jmol AstexViewer SPICE
CH	4 - 119	3KY9	A	4 - 119	Jmol AstexViewer SPICE
			B	4 - 119	Jmol AstexViewer SPICE
PH	403 - 504	3KY9	A	403 - 504	Jmol AstexViewer SPICE
			B	403 - 504	Jmol AstexViewer SPICE
RhoGEF	198 - 372	3KY9	A	198 - 372	Jmol AstexViewer SPICE
			B	198 - 372	Jmol AstexViewer SPICE
SH2	671 - 745	2LCT	A	671 - 745	Jmol AstexViewer SPICE
		2ROR	A	28 - 102	Jmol AstexViewer SPICE

Comments or questions on the site? Send a mail to pfam-help@sanger.ac.uk. Our cookie policy.
The Wellcome Trust

图4-23 人类原癌基因VAV的结构信息

在Jmol中查看C1_1的A链在PDB中的三维结构，如图4-24中紫色所示。

wellcome trust sanger institute

PDB entry 3KY9

Jmol

PDB			UniProt			Pfam family	Colour
Chain	Start	End	ID	Start	End		
A	198	372	VAV_HUMAN	198	372	RhoGEF (PF00621)	
A	516	568	VAV_HUMAN	516	568	C1_1 (PF00130)	
A	4	119	VAV_HUMAN	4	119	CH (PF00307)	
A	403	504	VAV_HUMAN	403	504	PH (PF00169)	
B	198	372	VAV_HUMAN	198	372	RhoGEF (PF00621)	

图4-24 人类原癌基因VAV的C1-1的A链的三维结构

(二) PIR数据库

PIR全称为The protein information resource，是一个集成了关于蛋白质功能预测数据的公共资源数据库，其目的是支持基因组/蛋白质组研究。PIR与MIPS(the Munich information center for protein sequences)、JIPID(the Japan international protein information dDatabase)合作，共同构成了PIR-国际蛋白质序列数据库(PSD) ——一个主要的已预测的蛋白质数据库，包括25 0 000个蛋白。为了提高蛋白质预测和实验数据之间的相互吻合程度，PIR建立了一套系统，允许研究者们递交、分类、提取文献信息。PIR提供了在超家族、域和模体水平上对蛋白的分类。PIR同时提供了蛋白的结构和功能信息，并给出了与其他40个数据库之间的相互参考。PIR还提供了一个非冗余的蛋白质数据库，包括从PIR-PSD、SWISS-PROT、TrEMBL、GenPept、RefSeq 、PDB收集来的约800 000条序列，对每条序列给出了一个符合的名称和相关文献。为了提高数据库的协同工作能力，PIR采用开发的数据库框架，利用XML技术进行数据发布。在PIR的站点上(http: //pir.georgetown.edu/)也提供了常规的生物信息学工具，以进行数据挖掘。

(三) 目前常用的蛋白质结构和功能数据库 (表4-9)

表4-9 蛋白质结构和功能关系数据库

数据库	结构信息	网址	功能信息
SignalP	信号肽	http: //www.cbs.dtu.dk/services/SignalP	蛋白质信号肽信息
ScanProsite	结合位点	http: //us.expasy.org/tools/scanprosite	检索Prosite数据库的快捷方式，提供结合位点描述信息
Pfam	结构域	http: //pfam.sanger.ac.uk/	结构域常用数据库，提供结构域功能描述
SMART	结构域	http: //smart.embl-heidelberg.de	结构域常用数据库，提供结构域功能描述
InterPro	结构域	http: //www.ebi.ac.uk/interpro/scan.html	结构域常用数据库，提供结构域功能描述
MATA	拓扑结构	http: //cubic.bioc.columbia.edu/predictprotein/submit_met.html	可自动链接到不同的拓扑结构分析程序
TMHMM	跨膜结构	http: //www.cbs.dtu.dk/services/TMHMM-2.0	常用的跨膜结构预测平台
PSORT	细胞定位	http: //psort.nibb.ac.jp/form2.html	查找细胞定位信号或基序
PDB	3D结构	http: //www.pdb.org	新发现蛋白质通常为阴性结构，但可与同源蛋白质进行结构比较
MIPS	物理结构互作	http: //www.mips.biochem.mpg.de/proj/yeast/tables/interaction	收集酵母中蛋白质相互作用
COG	同源性家族	http: //www.ncbi.nlm.nih.gov/COG	存储多物种同源蛋白质信息，蛋白质家族信息

三、从蛋白质结构推断其功能的方法与分析软件

蛋白质必须有特定的三维空间结构，才能表现其特定的生物功能。蛋白质的进化中保守的三维结构通常对应某些保守的精细生物化学功能，结构相似的蛋白质会有某些相似的精细生化功能。因此单纯依靠蛋白质序列相似性无法预测其功能。

早期基于结构预测蛋白质功能的方法是搜索与目的蛋白质结构相似的蛋白质，并将其功能赋给目的蛋白质，如DaLiLite、SSM、STRUCTAL、MultiProt、3Dcoffee等，这些方法将结构问题转化为序列问题，利用经典序列相似性算法来衡量结构相似性。虽然蛋白质三维结构对预测蛋白质功能很有意义，但这并不意味着知道蛋白质结构就一定知道其功能，这主要由于蛋白质功能依赖于其所处的细胞环境，而且蛋白质的折叠修饰极大地影响蛋白质功能，所以依据蛋白质结构预测其功能十分困难。目前还缺少仅依赖于蛋白质结构直接预测其功能的方法。一般的做法是通过识别蛋白质结构上的活性位点，结合区域或同源折叠关系为预测蛋白质功能提供线索。该方法很依赖蛋白质模型的可靠性，从头预测法很难满足蛋白质功能特征识别的需求，然而可以利用模糊功能结构（fuzzy functional forms，FFFS）来实现，即使用碳原子和侧链的中心位置来设计识别特定的结构模体进而预测功能的算法。

还有些利用其他途径来预测蛋白质功能的方法，如基于同源的进化分析方法、基于功能域的分析方法、基于基因表达簇的分析方法等。综合不同方法在特定生物学背景下解决结构比对的问题，有助于提高通过结构预测功能的精确性。如Michael、Edward等人都结合多种方法进行研究并得到相对理想的结果。

下面介绍一些现有的基于结构预测功能的方法及软件，这些方法分为四类：

1. 基于相似性的方法及软件　对于给定蛋白质结构，通过结构比对技术基于相似性方法（similarity-based approaches），来识别结构相似蛋白来预测功能。几种常用的基于相似性预测功能的方法（表4-10），表中前九个是两两比对算法，最后两个是多重比对算法。不过这些方法会受到缺少确定功能的蛋白质、结构域功能相似性之间存在差异等限制。

表4-10　几种常用基于相似性预测功能的方法

方法	参考文献
DaliLite	Holm and Park 2000
CE-MC	Shindyalov and Bourne 1998
SSAP	Oren go and Taylor 1996
SSM	Krissnel and Henrick 2004
STRUCTAL	Kolodny and linial 2004
LSQMAN	Kleywegt 1996
Proknow	Pal D and Eisenberg D 2005
VAST	Thompson KE et al.2009
FLORA	Redfern OC et al.2009
MultiProt	Shatsky et al.2004
3DCoffee	O’ Sullivan et al.2004

另外,基于FSSP数据库设计的PHUNCTIONER方法,通过识别每个蛋白质中GO分类特异的结构位点,对蛋白质残基保守性进行Z值(Z-score)计算来预测功能,精确度达到75%~90%。ROC分析结果显示其精确性和灵敏度比基于简单序列的方法更高。

还有,基于多维标度技术(MDS)的方法,依据具有相似功能的蛋白质相互近邻的推断,将蛋白质定位在蛋白质结构空间图(SSM)中,使用DaliLite方法进行打分判断蛋白质的GO分类。同样通过ROC分析显示,该方法同样好于基于简单序列相似性的方法。

【例4-5】用ProKnow预测人类乙醛脱氢酶的功能

(1)打开ProKnow(http://services.mbi.ucla.edu/proknow/)(图4-25)。

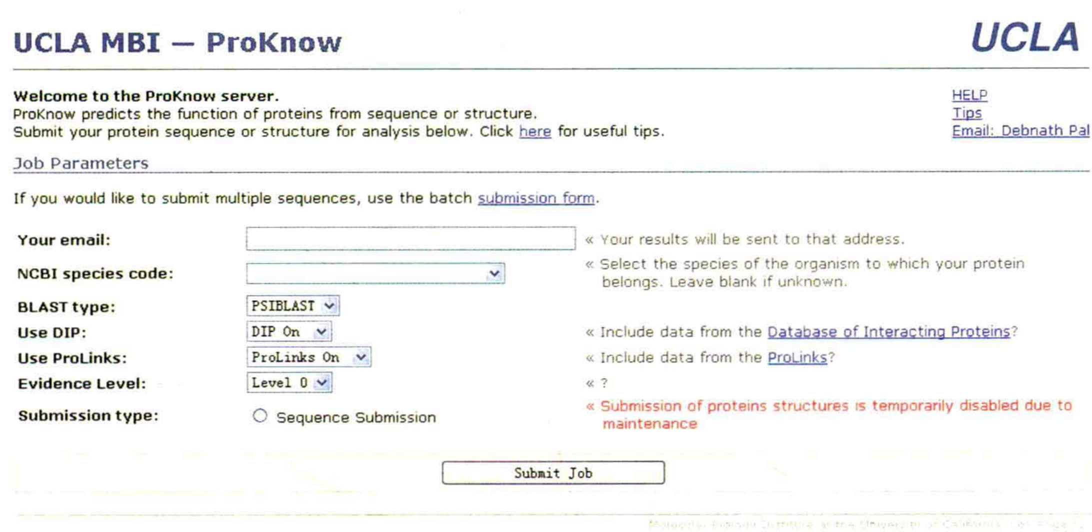


图4-25 ProKnow主界面

(2)在Your email处输入你的邮箱地址以便查询结果返回;选择物种为人类H.sapiens(9606);选中"Sequence Submission";输入人类乙醛脱氢酶的序列(FASTA格式);点击"Submit Job"进行搜索(图4-26)。

(3)等待一段时间后可得到结果(图4-27):预测的乙醛脱氢酶可能的功能为"aldehyde dehydrogenase [NAD(P)+] activity"、"oxidoreductase activity"、"phosphopyruvate hydratase activity"、"aldehyde metabolic process"、"negative regulation of metabolic process" 和"negative regulation of glycolysis"。

2. 基于三维基序的方法 这一类方法试图识别三维基序(3-dimensional motif-based),即保守亚结构,建立蛋白质功能和结构基序的关系映射来预测蛋白质功能。通过结构比对的保守性分析策略,可以有效地预测蛋白质功能。基于这种策略有许多方法和软件(表4-11)。

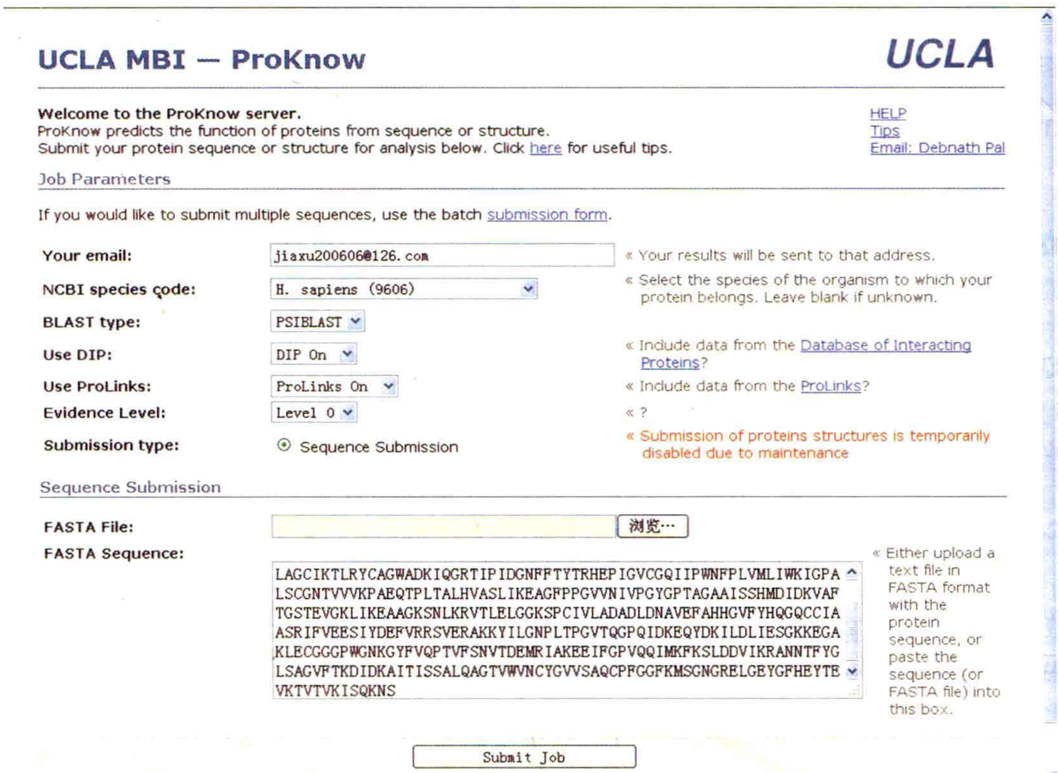
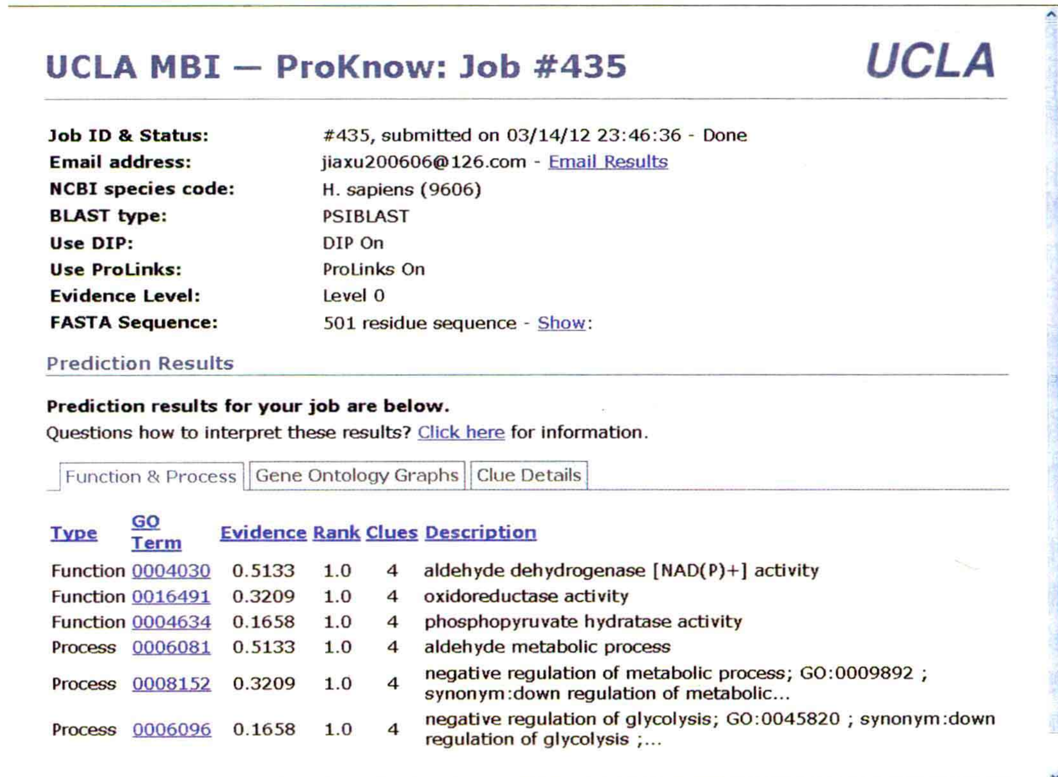


图4-26　在ProKnow主界面中输入要搜索的蛋白质序列及其他参数



Type	GO Term	Evidence	Rank	Clues	Description
Function	0004030	0.5133	1.0	4	aldehyde dehydrogenase [NAD(P)+] activity
Function	0016491	0.3209	1.0	4	oxidoreductase activity
Function	0004634	0.1658	1.0	4	phosphopyruvate hydratase activity
Process	0006081	0.5133	1.0	4	aldehyde metabolic process
Process	0008152	0.3209	1.0	4	negative regulation of metabolic process; GO:0009892 ; synonym:down regulation of metabolic...
Process	0006096	0.1658	1.0	4	negative regulation of glycolysis; GO:0045820 ; synonym:down regulation of glycolysis ;...

图4-27　ProKnow结果显示图

表4-11 几种常用的基于三维基序预测功能的方法

方法	描述
SITE算法	数据库存储了酶活性位点保守基序信息，通过位点匹配程序寻找关键的功能位点残基作为保守残基；但是在处理高度同源蛋白时会遇到其残基不保守的情况，因此需仔细分析这些信息预测功能
TESS算法	采用几何散列算法，通过模板分析和重叠从蛋白质的高级结构中寻找保守的必需残基，将未知蛋白与必需残基进行比对预测其功能，但该方法对高级结构预测精度要求较高
FFFs（模糊功能形态）	该方法基于几何形状、残基一致性和蛋白活性位点的证实对蛋白质进行三维描述来预测蛋白质功能
SPASM	同时用主链α碳原子和侧链基团作为分析对象，寻找并列的保守残基，并用于搜索结构数据库中能匹配的已知功能蛋白进而预测功能
FCANAL	Fast Calculable Protein Function ANALyzer，通过定义重要功能残基构建kernel功能位点，对其他残基构建相似性矩阵，进而预测功能
ProFunc server	对查询蛋白质结构，包括序列和结构motif搜索、活性位点识别和全局折叠比较，进行基于结构和序列的功能预测

还有一些数据库可用于识别蛋白结构域及新蛋白功能预测，如PROCAT、PROSITE、PRINTS、SMoS和DSMP等。

3. 基于表面的方法　一个蛋白质结构被定义为一个由三组坐标组成的坐标组，每组坐标表示对应氨基酸的空间位置，这表示分子内的相互作用会影响氨基酸水平或原子水平上特定的生物功能。而蛋白相互作用通常由于分子表面互补性而发生，因此通过蛋白质表面结构的信息来预测功能的方法被提出。

常用的方法是用图论技术来解决表面匹配问题，将来自PDB数据库的蛋白质结构信息用MSP算法分析其静电潜能和疏水性，进而分析其生化功能的静电表面（eF-site）推测功能。另一种方法由Binkowski等提出，基于蛋白质表面模型分析溶剂或配基与蛋白质的关系，推测蛋白质功能。他们认为溶剂或配基可以帮助蛋白质发挥功能，然后用Edelsbrunner方法分析蛋白质结构，依据pvSOAR数据库进行预测估值。另外SURFACE数据库也可对蛋白质进行局部表面特征（local surface patterns，clefts）模式识别，数据库选用SURFNET算法识别clefts，使用PROSITE数据库进行GO功能注释，结合RMSD和PAM矩阵进行测量打分、预测功能，精确性可达90%。

4. 基于机器学习的方法　利用有效的分类方法如SVM和KNN等，筛选最相关的结构特征中识别最适合的功能分类。基于机器学习的方法在功能预测上有着很大的成绩，如通过数据挖掘和机器学习方法的研究，分析两个数据对象之间的相似性的可变模型来预测功能。比较具有代表性的方法有三种：①$K_{pattern_Sim}$（S，T），基于相关氨基酸对定义两个亚结构之间的相似性；②K_{Redox_Func}（S，T），基于蛋白质基序CxxC定义巯基化合物/二硫化合物和氧化还原酶蛋白的功能相似性；③K_{3Dball}（P_1，P_2），前面提到前两种算法都是依据氨基酸位点定义相似性，而该算法是由给定半径的球形内的一组氨基酸定义的。利用核函数构建分类器，主要有K-NN（K-nearest neighbor）和SVM两种分类器，有实验表明K-NN比SVM具有更好的预测效

果。但是基于机器学习方法的障碍是缺少可信的蛋白质结构信息,因此随着对蛋白质领域的进一步研究,这种方法会有更好的效果。

四、蛋白质相互作用与蛋白质功能

生物系统的功能是由分子之间的相互作用而不是单个分子决定的。功能基因组学的一个主要任务就是理解蛋白质相互作用规律,这一工作对阐明蛋白质功能并进一步理解生物学过程有重要意义。已有大量高通量的实验方法可用于探测蛋白质相互作用。用实验方法研究蛋白质相互作用既费时又费力,这使得目前拥有的相互作用数据只占全部数据的一小部分,由此构建的相互作用网络也极不完整。作为一种补充,发展理论预测方法就尤为迫切。从蛋白质结构出发预测蛋白质-蛋白质相互作用(蛋白质对接)能够揭示它们的功能机制以及在细胞中起到的作用。现有的预测方法有基于DNA序列的基因近邻法、基因融合法、种系轮廓发生法,也有基于蛋白质一级结构的方法和基于蛋白质三级结构的方法。主要方法如下:

1. 同源建模　将已知三维结构的蛋白质复合物相互作用的信息应用到与组成该复合物的氨基酸序列的同源蛋白质间。Aloy等建立了这个方法,他们通过评估同源蛋白家族中已知3D结构的复合物的接触点特征,给出判断,并最终用实验的方法进行验证,准确率达到了65%。

2. 计算机模拟分子对接　早期的分子对接方法用分子力学方法或者量子化学方法计算小分子之间的识别,在一些分子模拟软件包中也含有分子对接的模块。但是由于算法和计算机处理能力的限制,早期的对接方法较难处理含有大分子的分子对接过程。1995年由Accelrys公司开发的计算化学软件Affinity上市,这是第一个可以进行有大分子参与的分子对接过程的商业化分子对接软件。此后,商业化和免费的分子对接软件层出不穷。现在应用中的分子对接软件涵盖了刚性对接、半柔性对接、柔性对接等各种对接方法,在能量优化方面则使用了人工神经网络、遗传算法、模拟退火、禁忌搜索、局部搜索等各种方法。目前的分子对接方法是研究小分子与大分子相互作用模式、生物大分子间识别、分子自组装、超分子结构等课题的常用方法之一。

3. 基于二级结构　统计计算蛋白复合体相互作用结合区域内不同二级结构及超二级结构出现的频次,所统计的二级结构主要分为三类: α螺旋、β折叠、无规则卷曲; 超二级结构是在二级结构基础之上的结构类型,工作中主要采用四种分类类型: α拐角、α发夹、β发夹、拱形结构。在统计数据的基础上计算不同结构类型出现在相互作用结合区的相对倾向值,并以此对蛋白质亚基对进行打分,将打分分值作为特征值输入支持向量机构建模型,对蛋白质亚基相互作用进行预测。

4. 基于结构域　蛋白质之间通过特异性的结合才能够发生相互作用,而这些结合部位就是结构域,因此,一种现在比较流行的思想是认为蛋白质间的相互作用是由蛋白质结构域之间的相互作用导致的。有学者提出假设,结构域组合间的相互作用是蛋白质相互作用中的基本单元。由于考虑到了结构域间相互作用导致蛋白质间相互作用的所有可能方式,预测效果有了显著提高。但是该模型仍然存在缺陷,即为了获取待预测样本中所有结构域组合对相互作用的概率,需要大量蛋白质相互作用数据以及相应蛋白质的结构域注释信息,而

训练样本的不足将导致无法获取所有结构域组合对的统计特征，从而只能评价部分蛋白质对的相互作用关系。使用支持向量机分析结构域组合对序列的氨基酸理化性质得到其序列特征值，同时采用统计分析的方法获取其频率特征值，最后通过融合上述两种特征估计该结构域组合间发生相互作用的可能性，并以此预测蛋白质间的相互作用关系。

从已知的蛋白质相互作用中预测有哪些结构域间存在相互作用，然后再利用预测结果对待测蛋白质对之间的相互作用情况进行判别。Gomez等首先提出来了一个简单的吸引模型，并定义了蛋白质相互作用的概率与结构域相互作用的概率关系模型。随后，他们又对该模型进行了改进，提出来了AR模型，在Pfam数据库上得到了较好的测试仪结果。

Deng等使用极大似然估计MLE的方法来寻找结构域间的相互作用，并反过来预测蛋白质间的相互作用，该方法能够对不完整的数据集和数据集中的错误进行有效的处理。Liu等将MLE方法进行了改进，使其能够利用多个物种中的蛋白质相互作用数据，提高了结构域相互作用预测的性能。Hayashida等将MLE中的问题进行了重新定义，使用线性规划的方法，通过最小化训练集中观测的相互作用与预测的相互作用之间的误差来对结构域间的相互作用进行预测，得到了较好的预测效果。Guimaraes等同样使用了线性规划方法，但其假定蛋白质间的相互作用符合简约原则，并以此建立线性规划的约束条件。

Huang等将结构域相互作用问题转化为集合覆盖问题，并提出了相应解决方案。Singhal等使用遗传算法，利用参数优化的思想计算结构域间的相互作用概率。Riley等使用结构域对排除分析方法，从多个物种的蛋白质相互作用数据中寻找潜在的结构域相互作用对。

还有一种从蛋白质域信息出发，分析互作蛋白质对和不互作蛋白质对各自的特征模式，基于极大熵聚类算法分析并预测蛋白质相互作用的方法，并采用von Mering数据集和DIP数据库中的数据测试了该方法，其预测的敏感性和特异性分别为92%和94%。基于上述方法，开发了网页工具用于预测蛋白质对的相互作用（http: //219.217.238.183 : 7001/prepi/index.jsp）。

5. 基于结构特征的方法　此方法注重蛋白质间的物理相互作用，包括相互作用界面（interface）及相互作用位点（interaction sites）的预测.蛋白质相互作用界面指的是两条以非共价键形式（non-covalent）结合的多肽链之间的共同区域，主要由对蛋白质结合起关键作用的、进化速率低于蛋白质表面其他部分的残基所组成。Aytuna等人在待预测数据集中寻找与已知蛋白质相互作用界面的互补对（complementary pairs）结构相类似的表面区域，通过蛋白质三级结构比对以及热点（hot spots），一种突变产生高能量且对蛋白质间相互作用的亲和性、稳定性起重要作用的残基匹配的方法推理预测PPI。例如，已知界面A与B存在相互作用，而表面区域a，b分别与A，B的结合位点在结构上具有相似性，从而推理a与b也存在相互作用，结果表明该方法具有较高的可信度。然而，三维结构已知的蛋白质数量的有限性限制了该方法的应用。Nussinov研究小组先后于1996年和2004年构造了蛋白质相互作用界面的非冗余数据集，界面数量由最初的351增加至3799，提高了该种方法的预测精度。此外，Gomez、Deng等人通过观察两蛋白质所含有的结构域之间是否存在吸引或排斥作用来预测PPI。他们利用Pfam数据库提供的域信息，分别采用AM（association method）和MLE（maximum likelihood estimation）方法，估算相互作用蛋白质对所含的结构域，并计算结构域信息的显著性和概率值，以此作为PPI存在的标签——这在本质上是一个机器学习过程。预测得到的相互作用蛋白质对其编码基因表达谱具有强相关性，证明了该方法的有效性。

6. 常用的在线分析工具

（1）PREPPI（http: //bhapp.c2b2.columbia.edu/PREPPI/）：整合结构与非结构信息预测蛋白质互作的网络工具。

（2）3DID（http: //3did.irbbarcelona.org/）：高通量三维结构已知蛋白质的结构域互作数据库。

（3）cons-PPISP（http: //pipe.scs.fsu.edu/ppisp.html）：用神经网络方法预测蛋白质互作的网站。输入一个蛋白质的结构，可以预测其与另一个蛋白质结合的位点残基。

（4）InterPreTS（http: //www.russelllab.org/cgi-bin/tools/interprets.pl）：通过三维结构预测互作的在线工具。

【例4-6】用PREPPI预测蛋白质互作

（1）打开网址:（http: //bhapp.c2b2.columbia.edu/PREPPI/）（图4-28）。

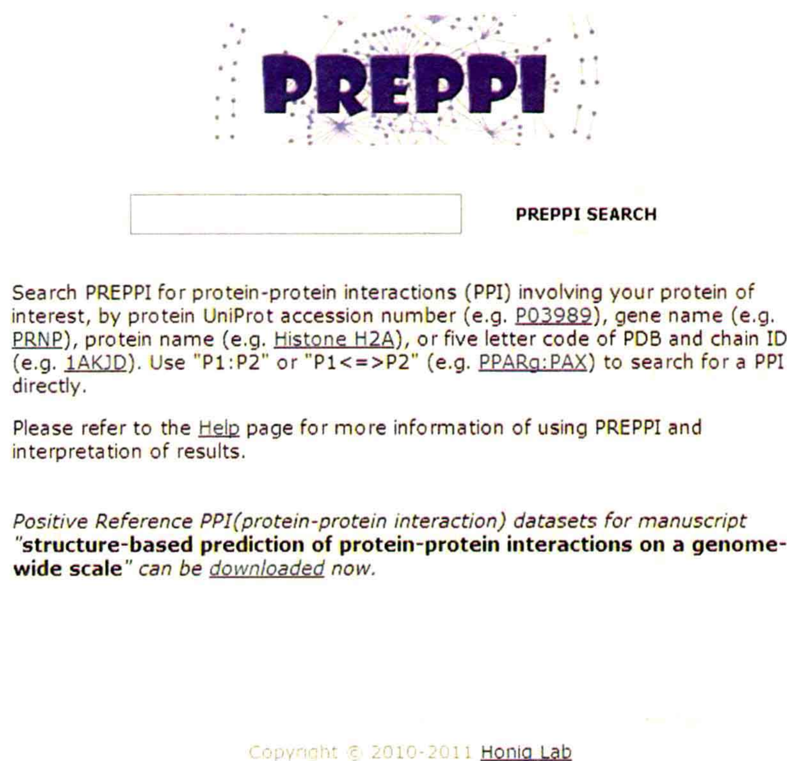


图4-28　PREPPI主页

（2）在搜索框中输入要查询蛋白质的名字、Uniprot ID或对应基因的名字，如：输入HLAB基因对应的蛋白质HLA I型组织相容性抗原的Uniprot ID“P03989”，点击“SEARCH”，可得到与其互作蛋白质的信息列表，结果如下（图4-29）。

结果包括所查询蛋白质HLA Ⅰ型组织相容性抗原的Uniprot ID、编码蛋白质的基因名、蛋白质名、蛋白质功能信息，同时还给出了被预测的与其互作的蛋白质的一些统计信息如：预测的高置信度的互作蛋白质125个（得分大于0.5）、所有预测的互作蛋白质298个（得分大于0.1），在数据库中存在的互作蛋白质12个。结果列表中的每一行是被预测为与HLA Ⅰ型组织相容性抗原互作的各蛋白质的信息，“Prediction code”列标明了确定该互作的信息来源，



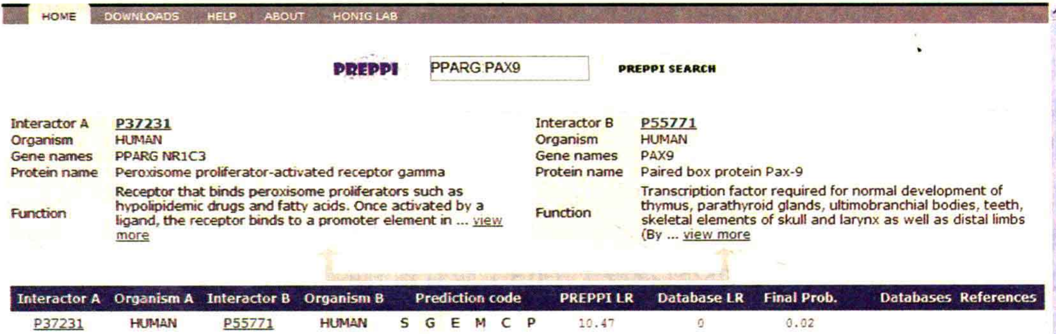
Interactor	Gene names	Organism	Prediction code	PREPPI LR	Database LR	Final prob.	Databases	References
P61769	B2M CDABP0092 HDCMA22P	human	S G E M C P	21096	4625.64	1.00	MINT	[1,2]
P01732	CD8A MAL	human	S G E M C P	9455.46	957.82	1.00		[1]
Q8NHL6	LILRB1 ILT2 LIR1 MIR7	human	S G E M C P	2767.55	957.82	1.00		[1]
Q8N423	LILRB2 ILT4 LIR2 MIR10	human	S G E M C P	2293.31	957.82	1.00		[1]
Q95460	MR1	human	S G E M C P	32199.5	NA	0.98		
Q30201	HFE HLAH	human	S G E M C P	7463.8	NA	0.93		
P30491	HLA-B HLAB	human	S G E M C P	7328.09	NA	0.92		
P30460	HLA-B HLAB	human	S G E M C P	7328.09	NA	0.92		
P30481	HLA-B HLAB	human	S G E M C P	7328.09	NA	0.92		
P30483	HLA-B HLAB	human	S G E M C P	7328.09	NA	0.92		
P30495	HLA-B HLAB	human	S G E M C P	7328.09	NA	0.92		
P30490	HLA-B HLAB	human	S G E M C P	7328.09	NA	0.92		
Q29718	HLA-B HLAB	human	S G E M C P	7328.09	NA	0.92		
P30685	HLA-B HLAB	human	S G E M C P	7328.09	NA	0.92		
P30475	HLA-B HLAB	human	S G E M C P	7328.09	NA	0.92		
P30479	HLA-B HLAB	human	S G E M C P	7328.09	NA	0.92		
P30484	HLA-B HLAB	human	S G E M C P	7328.09	NA	0.92		
P30462	HLA-B HLAB	human	S G E M C P	7328.09	NA	0.92		
P30493	HLA-B HLAB CDABP0067	human	S G E M C P	7328.09	NA	0.92		
Q29940	HLA-B HLAB	human	S G E M C P	7328.09	NA	0.92		
P18464	HLA-B HLAB	human	S G E M C P	7328.09	NA	0.92		

图4-29 查询蛋白质P03989的预测蛋白质互作结果

其中S表示结构信息，G表示功能信息，C表示共表达，P表示系统发育，E和M只在酵母数据中有，分别表示蛋白质必要性和MIPS信息。字母上的颜色表明其贡献程度，颜色越深贡献越大。"PREPPI LR"列表示用贝叶斯网络得到的整合"Prediction code"列中不同得分的计算预测LR。"database LR"表示实验得到的互作在各数据库中的整合LR。将LR值标准化得到"Final Prob."列中的值。最后两列表明储存该互作的数据库及证实文献。由结果中可以观察到，基于结构信息或以结构信息为主预测的互作蛋白质具有较高的置信度，并且其互作关系在现有数据库和文献中可得到证实，如结果图中列表显示的前4个蛋白质。

（3）可直接输入"蛋白质1：蛋白质2"或"蛋白质1<=>蛋白质2"查询蛋白质1与蛋白质2的互作情况。

例如，在搜索框中输入一对蛋白质过氧化物酶体增殖物活性受体和配对蛋白"PPARG:PAX9"，结果见图4-30。



Interactor A	Organism A	Interactor B	Organism B	Prediction code	PREPPI LR	Database LR	Final Prob.	Databases	References
P37231	HUMAN	P55771	HUMAN	S G E M C P	10.47	0	0.02		

图4-30 查询蛋白质对PPARG和PAX9互作结果

图4-30中分别列出了两个互作子蛋白质的Uniprot ID、编码蛋白质的基因名、蛋白质名和蛋白质的功能信息。表中各列意义如上所述。结果表明蛋白质PPARG和PAX9,没有基于结构的互作信息,只有基于功能互作的信息。其预测得分为0.02,这表明它们互作的可能性不大。

第五节

蛋白质的结构异常与疾病

Section 5 Protein Structure and Diseases

当蛋白质保守位点发生性质截然相反的突变时，蛋白质的高级结构可能被显著改变而影响其功能。另外，蛋白质序列不变而高级结构发生显著改变，例如变性（denaturation）或错误折叠（misfolding），也会造成蛋白质功能的显著改变，特殊情况下就会造成病理生理现象。

一、蛋白质序列变化引发疾病

分子病（molecular disease）是1949年由Pauling提出的，现已发现上百种。是指因某种蛋白质分子一级结构中的氨基酸残基序列与正常有所不同而发生的遗传病。如：镰状细胞贫血症（sickle-cell anemia）是一种常染色体隐性遗传疾病，患者的红细胞在缺氧状态下变成镰刀形。起因是体内合成血红蛋白的基因发生异常，使人血红蛋白β亚基的缬氨酸被谷氨酸所取代，只是一个氨基酸之差，则使患者的红细胞在缺氧状态下变成镰刀形，异常血红蛋白（HbS）从球状变为纤维状，而且易于在红细胞中析出。在一段时间内，此纤维状的HbS当氧分压高时，仍能恢复球状，但在氧分压降低时HbS又呈纤维状，这样几经恢复，使红细胞变得很脆弱，极易碎裂而发生溶血性贫血。当个体携带两个突变的β亚基基因时，会患镰状细胞贫血症。单一拷贝会引起镰状细胞特征，但通常无症状表现。编码β亚基的基因定位于染色体11-a区，含有许多β球蛋白基因簇，此区的多态性与疾病的严重程度有直接的关系。

二、蛋白质折叠错误引发疾病

蛋白质构型的改变是蛋白质错误折叠的主要原因。一般而言，天然构象主要由α-螺旋和无规卷曲组成，而错误折叠的构象富含β-折叠结构。例如：亨廷顿舞蹈病的发病机制主要与多聚谷氨酰胺的延长有关，当谷氨酰胺重复序列增长时，可促使蛋白质构型从随机缠绕（random coil）向β-折叠转化。聚合体是由反向排列的β-折叠不断增加形成的，α-螺旋/β-折叠的结构转换导致疏水基团暴露而亲水基团埋在蛋白质内部，引起蛋白质分子之间形成交叉的β-折叠结构。β-折叠之间由侧链和主链中的氢键将其连接在一起，多聚谷氨酰胺发生交联，溶解度降低，最后在胞质中形成聚合体或在核内形成包涵体。Htt聚集体具有细

胞毒性，可封闭转录因子，抑制泛素-蛋白酶体系统，导致神经元死亡。

朊病毒病是一种由朊病毒引起的与痴呆相关的神经退行性疾病，可在人和动物间传播。朊蛋白（prion protein，PrP）是一种能够在正常的哺乳动物的神经细胞表达的蛋白，人的PrP基因位于20号染色体，编码产生的蛋白称为PrPC，由GPI锚固定在细胞表面。在基因突变或环境变化或感染Scrapie等条件下，蛋白质的组成氨基酸顺序不变，但它的空间结构可发生变化，螺旋结构减少，β片层结构增加，称之为scrapie associated prion protein（PrPSc），其性质也随之发生变化，有细胞毒作用，可引起神经变性、胶质细胞增生和细胞外淀粉样沉积等病变。其变化形式为：PrPS以α螺旋结构为主，β折叠仅占3%。PrPSC中β折叠占43%，易于聚集，形成具有细胞毒性的高分子量的不溶性复合物的沉积而引起病变（图4-31）。

Garrett & Grisham: Biochemistry, 2/e
Unnumbered Figure p.979

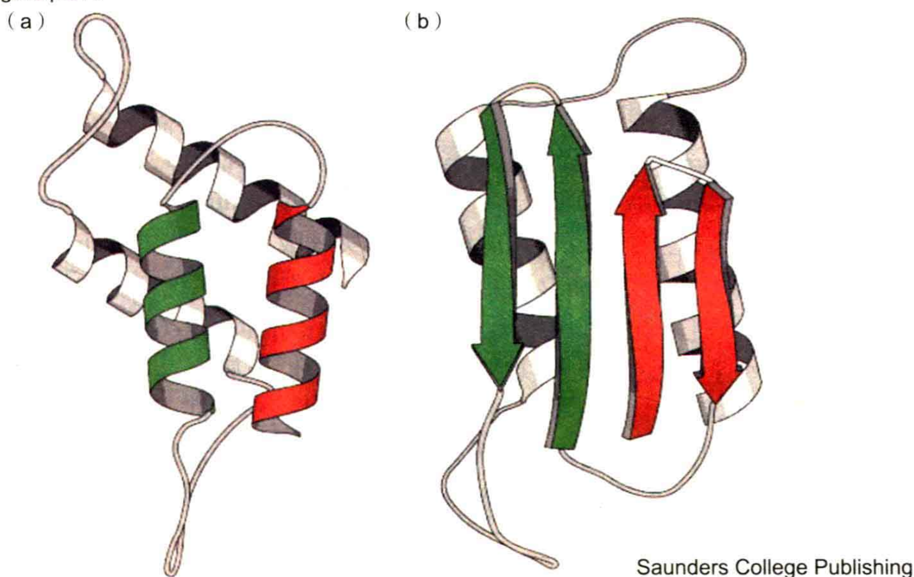


图4-31　正常的Prion蛋白含有大量的α螺旋；病变的Prion蛋白含有更多的β折叠

引自于：Garrett & Grisham, Biochemistry, 2nd ed., 1998：Schematic view of the two structures of a protein involved in the neurodegenerative disease of sheep, scrapie.

阿尔茨海默病（Alzheimer's disease）患者体内特别是脑内的淀粉样蛋白浓度显著升高，特别是42个氨基酸残基的片段含量不成比例的升高，β折叠增加，易于积聚而形成淀粉样斑块。神经纤维缠结主要由高度磷酸化的微管相关的tau蛋白异常折叠聚集而成。通过对错误折叠机制的研究，可以明确病理机制，使临床方案更具有效性和精确性，对遗传学和医学研究具有一定意义。

三、疾病过程中蛋白质的相互作用

蛋白质的高级结构决定了其在生物体内的功能，多个蛋白质发挥作用时常需要与其他蛋白质协同作用，不同蛋白质之间形成复合体（complex）。每个蛋白质可以看成复合体的一个亚基（subunit），亚基间相互作用，形成紧密的复合体结构或共同组成复合体的活性中心。

当蛋白质序列变异或发生错误折叠等结构突变时，正常的蛋白质结构互作缺失，引起特定的功能或表型异常，就会引发疾病。

大量关于p53各方面功能及其在人类肿瘤中表达的突变蛋白质的信息，阐明了p53在癌症发病机制中的重要作用。p53是一种四聚体，单独一个亚基的突变就会影响整个复合物的活性，导致DNA结合活性的减弱。目前已经发现超过14 000个与肿瘤相关的p53突变。但人类腺病毒的E6蛋白可通过与p53的Arg72高亲和性地相互作用，同样可使p53蛋白质失去功能，阻止它对损伤DNA的修复，从而导致子宫癌的发生。

通过研究霍乱毒素的结构及功能，发现其毒素是87kDa的六聚体（亚基组成为AB_5）蛋白质，它通过B亚基结合在GM1的神经节苷酶上，并将A亚基通过受体介导的胞吞作用转运入膜内。在细胞内，二硫键发生还原性断裂，从A亚基上释放一个包含195个氨基酸残基的片段，此片段催化ADP-核糖由NDP^+转移至异源三聚体G_S蛋白G_α亚基的Arg187侧链上，这种糖基化过程持续地激活腺苷酸环化酶同时抑制G_α的GTP酶活性，从而使细胞内cAMP水平剧烈增高，导致肠道细胞激活钠泵，分泌Na^+。为抵消氯化物，水与碳酸氢盐也被分泌出去，最终的网络效应导致大量水和电解质缺失，引发疾病，导致脱水症，最终可引起死亡。另外，禽流感病毒能否感染人类取决于它的血凝素（病毒表面的一种蛋白）是否能够与呼吸道多糖受体结合。研究人员借助NIGMS的一个专门数据库（consortium for functional glycomics），进行蛋白质与不同类型糖分子相互作用的研究。人类呼吸系统细胞中有alpha2-6类的多糖受体；禽类呼吸系统细胞中则是alpha2-3类多糖受体。人类呼吸系统细胞中的alpha2-6类多糖受体有两种形状，分别为伞形和圆锥形。病毒可与圆锥形alpha2-6受体结合，但人类呼吸道中这种受体远远小于伞形受体，所以感染能力差。因此，流感病毒如果要感染人类，必须与伞形的alpha2-6受体结合。可以寻找那些已经进化出与伞形alpha2-6受体结合的病毒，并针对其开发新疫苗，以便应对可能暴发的大规模流感。随着蛋白质精细结构的逐步解析，从蛋白质结构互作的角度来研究和探索复杂疾病的潜在发生机制，进而进行药物研发具有重要的意义。

第六节

应 用 实 例

Section 6 Application of Disease

一、蛋白质结构信息在类风湿性关节炎致病基因挖掘中的应用

类风湿性关节炎是一种复杂的炎症性疾病，主要与关节、自身免疫功能以及遗传因素有关。治疗类风湿性关节炎的重要挑战是找到一种有效的筛选方法，寻找到与已知疾病基因有相似结构和功能的候选风险基因，并利用它们开发新技术用于检测、诊断和治疗。蛋白质是生物体重要的组成部分，并参与几乎每一细胞过程。大部分蛋白质折叠成独特的结构，以便在不同功能集中的具体特性做出特定的贡献。致病蛋白质与致病基因往往是通过相似的序列和结构相关联的，所以候选基因可以通过序列及与已知致病基因相似的晶体结构筛选出来。本案例通过采用统计遗传学贝叶斯关联分析方法和模式识别的方法，针对已知致病基因与非致病基因编码蛋白质在序列和结构特征上的差异，来预测类风湿性关节炎的致病基因。并从家族功能特性、GO功能一致性和KEGG通路富集三方面对预测的致病基因进行评价，以期找出与类风湿性关节炎疾病发病机制密切关联的疾病基因。

1. 实验数据资源

从在线GAW16（http://www.gaworkshop.org/）中下载，868个类风湿性关节炎样本和1194个正常样本的SNP基因型频率数据。

人类基因的序列信息、位点信息来自NCBI（http://www.ncbi.nlm.nih.gov/）的基因组数据库。

疾病基因和疾病位点信息来自OMIM（http://www.ncbi.nlm.nih.gov/omim）数据库。

人类蛋白质的结构数据来自PDB（http://www.rcsb.org/pdb/home/home.do）数据库及targetDB 数据库（http://targetdb.pdb.org/）。

功能注释和功能鉴定分析资源主要采用PIRSF（http://pir.georgetown.edu/pirsf）中的功能分类、GO（http://www.geneontology.org/）注释体系和KEGG数据库（http://www.genome.jp/kegg/）。

2. 实验方法

（1）支持向量机（support vector machine，SVM）分类器中分类集合的构建：以下载的GAW16中检测的类风湿性关节炎的SNP群体数据作为研究对象2062个样本中含有433 766个SNP，对基因组层面的贝叶斯关联分析得到疾病与对照样本差异显著的SNP集，针对集合中的每一个SNP，根据其在NCBI数据库中对应染色体的物理位置，寻找在其上下游500kb范围内的基因，得到的基因集定义为候选疾病基因集合，共4402个基因，作为SVM分类器的检验

集；从OMIM疾病数据库得到的类风湿性关节炎相关疾病基因集，共335个基因，作为SVM分类器的阳性集（致病基因集）；假定在全基因组中，除去阳性集和检验集的集合为非致病集，共得到28874个基因，用该集合作为SVM分类器的阴性集。

（2）SVM分类器特征的确定：提取蛋白质的结构特征共28维（见表4-12），在提取特征的前20维为蛋白质一级结构特征，21~28维为蛋白质二级结构特征。用一级二级结构组合特征构建SVM分类器。

表4-12　蛋白质二级结构的二级特征维数，名称及其表达的意义

特征维数	特征名称	特征含义
1-20	C	Composition of the 20 amion acid residues
21	a	Cell length a in Angstroms
22	b	Cell length b in Angstroms
23	c	Cell length c in Angstroms
24	alpha	Cell angle alpha in degrees
25	beta	Cell angle beta in degrees
26	gamma	Cell angle gamma in degrees
27	helical	Percent of helical in protein sequence
28	Beta sheet	Percent of beta sheet in protein sequence

（3）分类器的确立：在PDB 数据库和 targetDB 数据库中分别针对候选基因集、非致病基因集及已知致病基因集，用网页文本挖掘的方法，筛选保留具有28维特征的蛋白质，得到已知致病集574个蛋白质，候选集2664个蛋白质，非致病集2385个蛋白质。

采用训练集：已知致病集（阳性集）和非致病集（阴性集），用5倍交叉证实来评估由一级二级组合特征构建的SVM分类器。通过对1000次分类结果的统计，发现应用组合特征的准确率为89%。最终选择此组合特征构建的SVM分类器，作为对检验集进行预测的分类器。

3. SVM分类器筛选结果分析及评价

（1）预测的致病基因：通过应用蛋白质一级结构和二级结构组合特征分类器对检验集的2664个蛋白质进行分类，预测得到候选致病蛋白质ID 944个，对应的候选致病基因495个。

分别对候选致病基因集（495个基因）和已知致病基因集（335个基因）进行蛋白质功能家族分类分析、GO功能节点富集分析和KEGG风险通路分析。选取与类风湿性关节炎已知致病基因共享至少有一个功能注释的候选基因作为评价后的预测的致病基因，即：对于495个候选致病基因进行评价，进一步确定了146个预测的致病基因。

（2）预测致病基因的功能分析：在PIR数据库中筛选已知致病基因集和预测致病基因集所属的家族，包括免疫球蛋白家族、Protein kinase domain 家族、SH3 domain 家族以及Ligand-binding domain of nuclear hormone receptor家族等。以免疫球蛋白家族为例（图4-32），可以发现14个已知致病基因与10个预测致病基因在该家族上富集，这体现了预测基因与已知基因在该家族功能上的一致性。

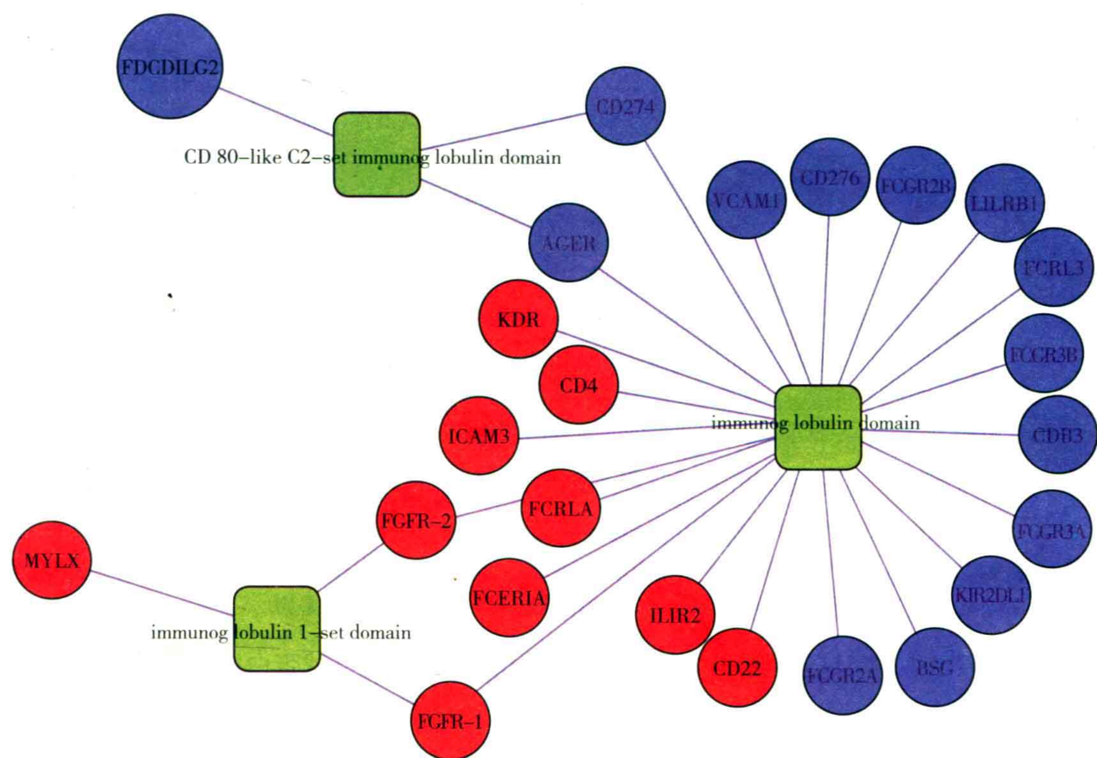
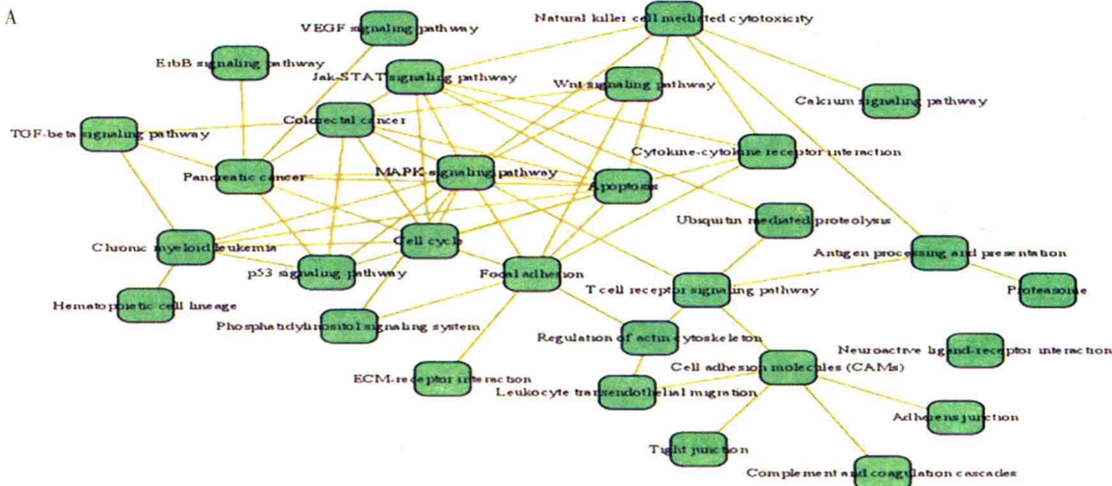


图4-32　**已知致病基因和预测的致病基因在Immunoglobulin家族上的富集示意图**

红色节点代表预测的致病基因，蓝色节点代表已知致病基因

应用GeneWebgestalt(http: //genereg.ornl.gov/webgestalt/)在线分析软件研究146个预测致病基因，发现其富集在信号转导、细胞过程的正向调节、免疫系统和免疫反应等功能上，与已知致病基因富集的GO功能节点一致。预测致病基因集和已知致病基因集共享相同的KEGG通路，如：重要的通路包括细胞因子与细胞因子受体互作通路、JAK－STAT信号通路，细胞黏附分子和MAPK信号通路等，而且这些富集的通路还相互紧密连接、相互作用，一起参与传递疾病风险（图4-33）。



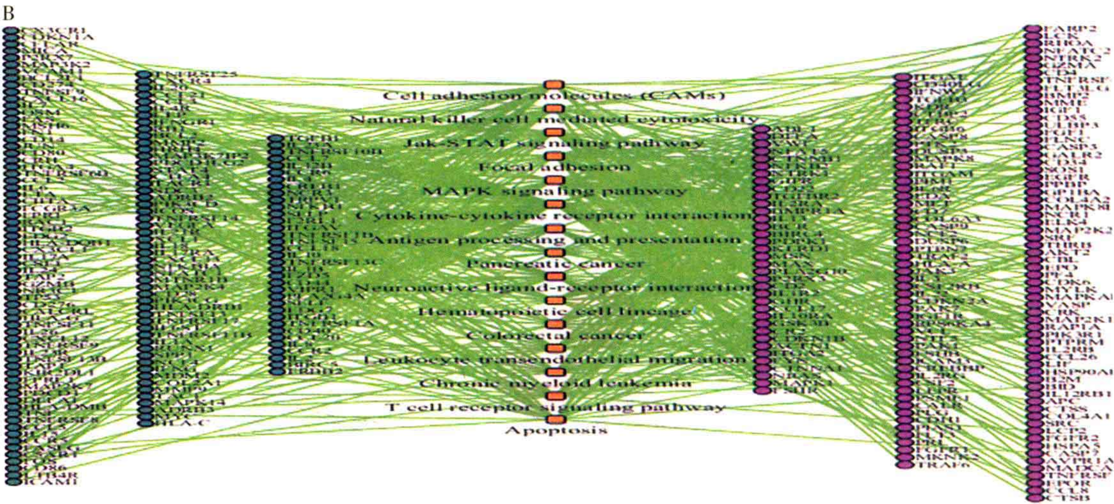


图4-33　通路互作与基因通路关系图

A.类风湿关节炎通路互作图，每一个绿色节点为一个通路，边代表互作；B.候选基因与相应通路之间的关系，橘色节点代表通路，蓝色节点代表已知致病基因，粉色节点代表预测致病基因

（3）预测致病基因与风湿性关节炎关系的文献证实：以免疫球蛋白家族为例，通过文献证实，发现免疫球蛋白家族、Protein kinase domain 家族、SH3 domain 家族以及Ligand-binding domain of nuclear hormone receptor家族的基因大都与类风湿性关节炎的发生发展相联系。最终，对基因评价后的146个预测的致病基因中，有41个基因得到了很好的文献证实。

二、蛋白质结构转换几率与疾病的发生

在生命体中，蛋白质几乎参与所有的生理过程。生化反应要求相关各功能团处于一定的距离内，因此蛋白质通常要折叠成特定的结构才能执行其功能。一旦蛋白质结构由正常状态转变为错误状态就可能导致疾病的发生。由于蛋白质所涉及生理过程的广泛性，错误折叠就成为了引发疾病的一种普遍因素。与蛋白质折叠异常有关的疾病，其触发致病结构变化的第一步，通常是在稳定的天然结构的重要区域发生错误折叠。这破坏了蛋白质的正常构象，显露出了先前被隐藏的聚集易发区，从而在错误折叠途径中导致了后续错误。第一阶段的位置可被认为是蛋白质的开关区，这些位置能作为药物靶点有助于阻碍异常折叠通路、并防止构象疾病的发生。

在正常的蛋白结构中，通常将构成蛋白质的短肽归纳为两大类别。一类以螺旋为标志性结构，另一类以β折叠为标志性结构，而某一短肽只属于两大类中的一种。为理解触发蛋白质构象疾病的相关机制，利用聚类分析方法，综合考虑影响蛋白质进化的突变、结构、力学属性等诸多因素，以一类短肽转换为另一类短肽的概率来判别该短肽触发致病性结构改变的能力。这种预测负责致病性结构变化起始交换区的算法称为构象疾病开关（CD_SWITCH）算法。

1. CD_SWITCH算法的实现

第一步：将一个查询的蛋白质视为连续的残基片段，以15-残基为一个片段，沿着蛋白质序列滑动一个15-残基窗口，查询蛋白质的每一个片段作为一个查询多肽。从同源多肽关

系图(GPR)中确定查询多肽的远程同系物(RHs)。

构建查询多肽的数据库$\tilde{\Gamma}$,数据库中的每一个节点设有一个条目标题和一个条目“索引”(Index)。对于每一个查询多肽i,会产生一套RHs$\left\{\Gamma_i^{RH_1},\Gamma_i^{RH_2},\Gamma_i^{RH_3},\cdots\right\}$:

Index, $\Gamma_i^{RH_1}$; $\text{Index}_{\text{homologue1}}$, $\text{Index}_{\text{homologue2}}$, $\text{Index}_{\text{homologue3}}$, …

Index, $\Gamma_i^{RH_2}$; $\text{Index}_{\text{homologue1}}$, $\text{Index}_{\text{homologue2}}$, $\text{Index}_{\text{homologue3}}$, …

Index, $\Gamma_i^{RH_3}$; $\text{Index}_{\text{homologue1}}$, $\text{Index}_{\text{homologue2}}$, $\text{Index}_{\text{homologue3}}$, …

第二步:评估螺旋圈区和折叠区之间交换的概率

对于每个查询多肽,用第一步得到的条目信息来评估查询多肽结构的改变的倾向。对于多肽的每一个15-残基,分别考虑前7-残基和后7-残基,只要7个残基中超过三个残基分布在螺旋构象中,则记为状态H;同理,只要7个残基中超过三个残基分布在折叠构象中,则记为状态E;否则记为状态C。随后,对于15-残基多肽定义了9个状态:HH、HC、CH、EE、CE、EC、HE、EH、CC。在图4-34中,HH+HC+CH的节点对应螺旋区相应的主体,EE+EC+CE的节点大部分在折叠区。因此,从多肽的二级结构中可以推断出在多肽空间中一个片段的位置。

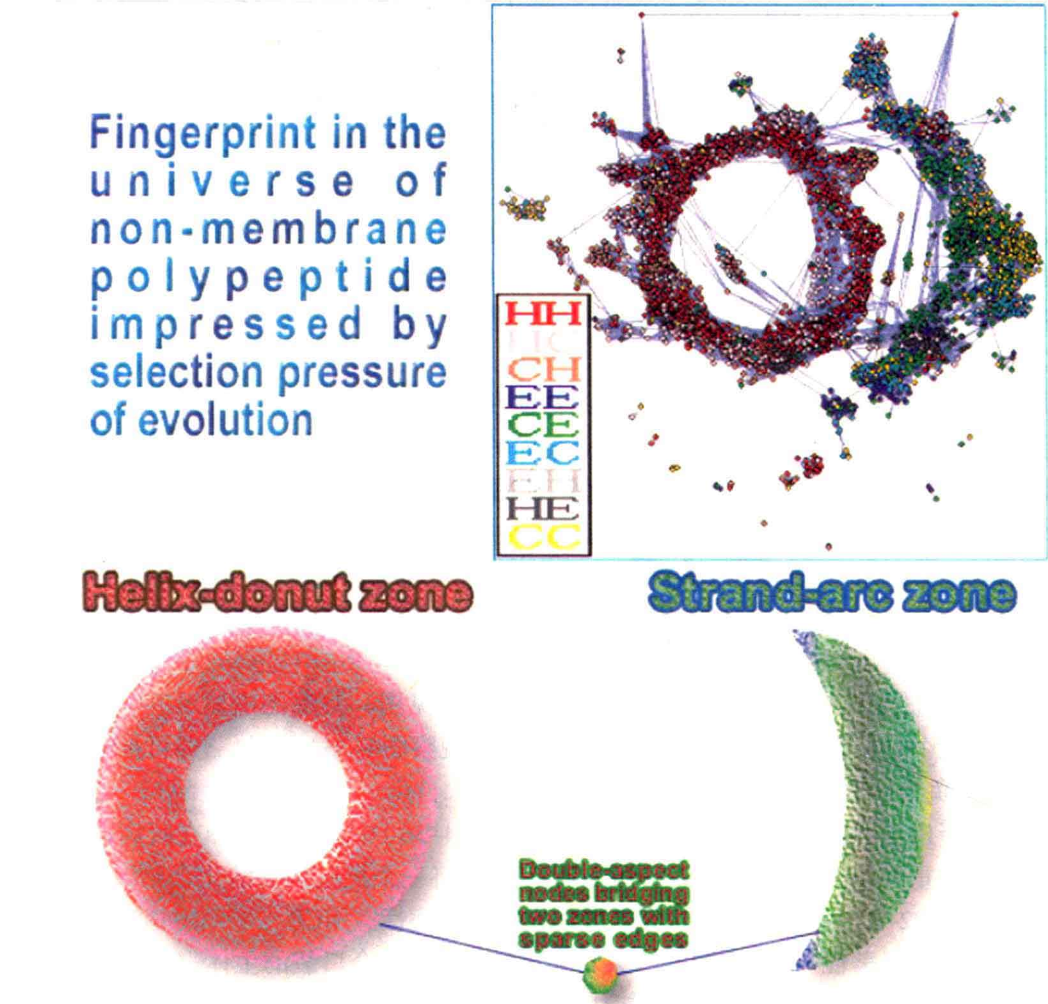


图4-34　同源拓扑特征示意图

对于查询多肽i，通过“索引”来评估它属于由点组成的螺旋圈区或折叠区的概率，设这个“索引”所组成的集合为{W_i}。用相关多肽的信息，评估查询多肽i在区域σ中的概率：$P_i(\sigma)=\sum_j^n \delta\left(\sigma,\sigma'\left(w_{ij}\right)\right)/n$，其中$n$代表{$W_i$}的规模，$w_{ij}$代表含有索引j的集合{$W_i$}中的一员，如果$w_{ij}$属于HH+HC+CH，则$\sigma'\left(w_{ij}\right)=0$，如果$w_{ij}$属于EE+EC+CE，则$\sigma'\left(w_{ij}\right)=1$；$\delta(x,y)$是阶跃函数，当x=y，则$\delta(x,y)=1$，否则$\delta(x,y)=0$；$\sigma=0$对应螺旋圈区，$\sigma=1$对应折叠区。每一个查询多肽可能位于螺旋圈区，或者折叠区，或者其他区。因为GRP中的大部分节点位于螺旋区，所以研究两个区之间的交换概率具有足够的代表性和准确性。通过结构提供的信息，对于查询多肽i，螺旋圈区和折叠区之间的交换概率用Q_i来评估：

（1）当查询多肽i由HH+HC+CH跳到折叠区时：

$Q_i=(1-P_{i-1}(0))P_i(1)(1-P_{i+1}(0))$，

（2）当查询多肽i由EE+EC+CE跳到螺旋圈区时：

$Q_i=(1-P_{i-1}(1))P_i(0)(1-P_{i+1}(1))$，

（3）当查询多肽属于其他区时：

$Q_i=0$

第三步：在正常条件下允许构象变化的过滤器

由于热运动，在室温不断变化下，一个蛋白质有一个正常的灵活结构，这种结构对于蛋白质分子生物功能是至关重要的。对于螺旋和β折叠构象，二级结构的这两个类型在正常条件下是可互换的，而像这种天然的二级结构的轻微扩张或收缩是不会引起疾病的。在一个查询多肽两侧同时扩展x个残基，考查在螺旋或折叠中是否存在一个放大的窗口。如果扩展后与扩展前的查询结果一致的多肽被过滤，其相应的Qi被设置为0。与构象疾病相关的查询蛋白质，在对应的结构变化中，含有高Qi值的多肽，被预测为开关位置。

2. 方法的评估与应用

用序列同源性低的蛋白证实了该方法的普适性，由于这个算法是基于远程同源性，对于相应蛋白质家族的所有成员的区域的确定是相同的，因此，对于输入蛋白质的结构没有严格的限制。并通过对几十种涉病蛋白质的检验，证明此方法对位于体液环境中的蛋白质或结构域均有效，适用于由蛋白质结构变化引起的各种疾病。

以人类朊病毒蛋白质（PrP）为例，首先，用人类PrP的每15-残基多肽来评估它们的交换概率。在这个分析中，每个残基用15个连续残基来表征，为了评估每个残基位置的显著性，用相应15个残基的最大概率对每个残基进行打分。图4-35A中显示的峰值代表发生在位置195残基处的交换概率。对于每个残基的交换概率见图4-35B，图中显示五个可能的开关位置中有四个是位于188~202区域内（红色区域）。观察图4-35A和图4-35B可推测出位置188~202应该负责构象变化的起源。

朊病毒蛋白质和抗朊病毒的化合物的结构图（图4-36）中显示了已知在朊病毒中阻碍致病性变化的重要结合位点。2007年，Kuwata等报道将一个抗朊病毒的化合物GN8结合到N159-V189-T192-K194-E196相应区域能抑制朊病毒致病性。用CD_SWITCH算法通过触发构象转换几率的显著性分析，预测的构象转换区域与GN8结合口袋的主体完全对应。所以对于朊病毒蛋白质（PrP），这一区域可作为此蛋白质相关结构病变的触发开关位置，因此验证了该算法预测结果的准确性。

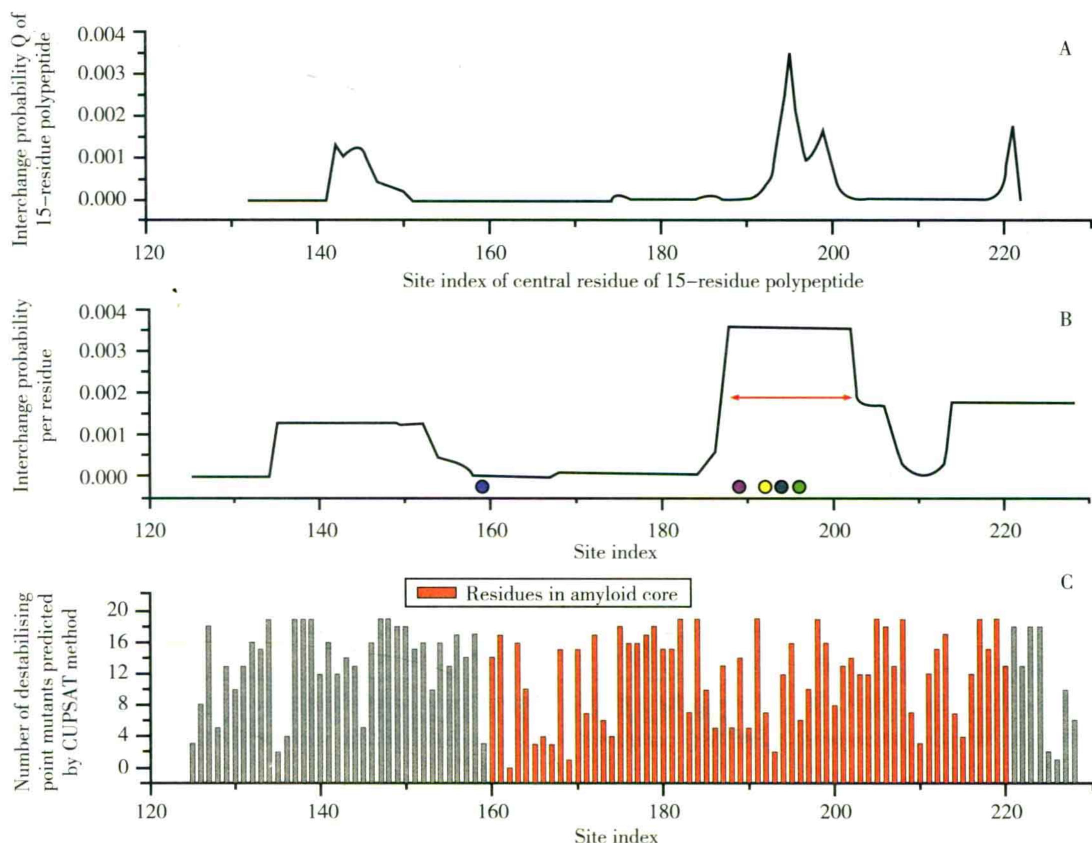


图4-35 人类朊病毒的结果(PDB ID1qm2_A,104个残基长度)示意图

A.每个15-残基片段索引的交换概率;B.每个残基位置的交换概率;C.在进化信息缺乏的情况下,对于预测朊病毒稳定性的显著性位置

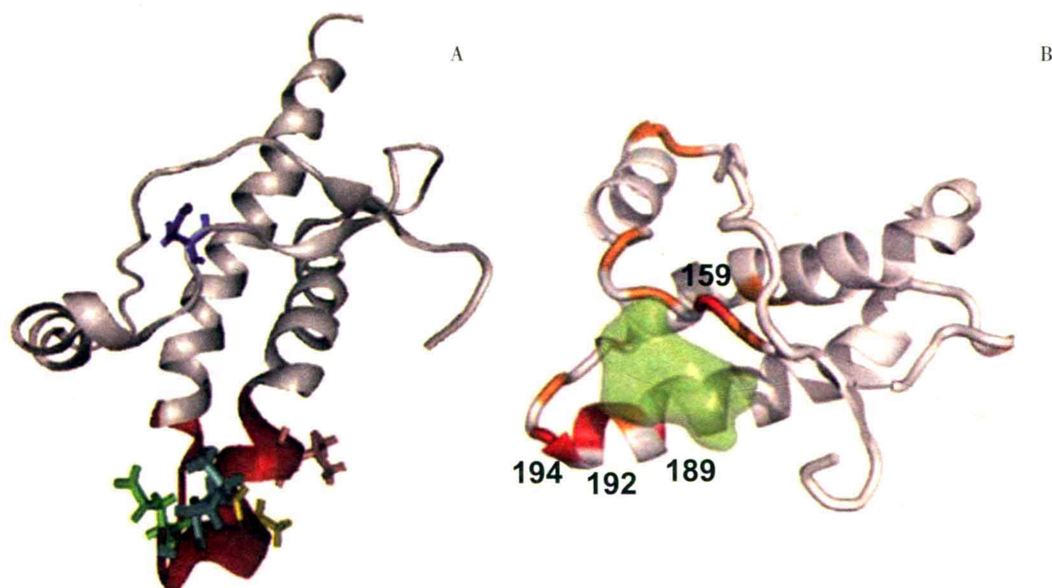


图4-36 朊病毒蛋白质和抗朊病毒的化合物的部分结构示意图

A.朊病毒的结构(蓝色159,紫红色189,黄色192,橄榄色194,绿色196);B.抗朊病毒化合物GN8的结合口袋

CD_SWITCH算法对病变敏感区域的预测精度可达94%,病变能力越低的蛋白发病后越致命。换句话说,蛋白质中的病变敏感区域可由理论迅速标定,大量用于测定此种区域的实验资源有望得以节省。研究表明对病态结构改变的研究并不拘泥于四十几种经典的蛋白质构象病,很多其他病理现象都可以据此进行分析,如高致病性H5N1型禽流感的高毒机制和2009甲型H1N1流感的种属跨越机制等。通过对错误折叠机制的研究,可以明确病理机制,使临床方案更具效率和精确性,对遗传学和医学研究具有一定意义,而从病态结构改变的角度进行病理机制的研究具有一定发展前景。

(陈丽娜 李琬 何月涵)

参考文献

1. Malik A. , Firoz A. , Jha V. , et al. *PROCARB: A Database of Known and Modelled Carbohydrate-Binding Protein Structures with Sequence-Based Prediction Tools.* Adv Bioinformatics, 2010 : 436036-44.
2. Pal D, Eisenberg D. . *Inference of protein function from protein structure.* Structure, 2005. 13(1): 121-30.
3. Aloy P. , Pichaud M. , Russell R. B. *Protein complexes: structure prediction challenges for the 21st century.* Curr Opin Struct Biol, 2005. 15(1): 15-22.
4. Wong E. , Thackray A. M. , Bujdoso R. , *Copper induces increased beta-sheet content in the scrapie-susceptible ovine prion protein PrPVRQ compared with the resistant allelic variant PrPARR.* Biochem J, 2004. 380(1): 273-82.
5. Chandrasekaran A. , Srinivasan A. , Raman R. , et al. , *Glycan topology determines human adaptation of avian H5N1 virus hemagglutinin.* Nat Biotechnol, 2008. 26(1): 107-13.
6. Zhang L. , Li W. , Song L. , et al. , *A towards-multidimensional screening approach to predict candidate genes of rheumatoid arthritis based on SNP, structural and functional annotations.* BMC Med Genomics, 2010. 3 : 38.
7. Liu X, Zhao Y. P, *Switch region for pathogenic structural change in conformational disease and its prediction.* PLoS One, 2010. 5(1): e8441.

第五章
CHAPTER 5

分子进化分析

MOLECULAR EVOLUTION ANALYSIS

进化是一种不断改进的过程,《物种起源》中这样描述:"每个生物每时每刻都在为生存进行反复的斗争,如果在复杂甚至多变的生存条件下该生物仍然能够不断改进自己,那么其将有较大的生存可能性并被自然选择所保留。根据严格的遗传法则,任何被自然选择保留下来的物种都倾向于繁殖其已经被改进的新的生命形式。"尽管自然选择在形态形成和行为进化方面似乎普遍存在,但在某些基因和基因组进化中所起的作用也有其他看法。分子进化的中性学说认为,种内和种间大多数可见差异不是自然选择,而是适合度很小的随机突变的固定所决定的。

人类基因组和多种生物基因组测序计划的完成,推动了分子进化的跨越式发展,基因表达和生物网络的进化等研究内容不断出现在最新的研究中,扩展了分子进化分析的研究范畴。许多研究者认为基因表达调控的差异可能对物种内和物种间的表型差异有重要的作用;基因的进化可能不是独立进行的,而是受到蛋白质互作或通路的限制,是一个协同进行的过程,这些研究拓展了分子进化的深层分子,此外多个基因共同进化或者以模块的形式研究进化关系,以及从整个网络的层面实现进化的研究。在本章下面的内容中,将对分子进化的基本知识和研究进程进行介绍。

第一节

系统发生分析与重建

Section 1 Phylogeny Reconstruction

一、核苷酸置换模型与氨基酸置换模型

(一) DNA序列进化分析

由于DNA序列包括多种不同类型的区域，如蛋白质编码区、非编码区、外显子、内含子、侧翼区、重复DNA序列和插入序列等。因此DNA序列的进化演变比蛋白质序列的演变更复杂。因此，弄清所研究的DNA类型和功能是十分重要的。即便我们单独考虑蛋白质编码区，密码子第一、二、三位的核苷酸替代样式也不尽相同。而且，某些区域比其他区域更易受到自然选择的影响，因此DNA不同区段呈现不同的进化模式。这里主要研究蛋白质编码区和RNA编码区，这些区域的进化相对简单，但通过它们来理解进化的一般规律极为重要。

1. 两个序列间的核苷酸差异　同一祖先序列传衍的两条后裔序列，它们的核苷酸差异随时间增长而增加。一个简便的描述序列分歧大小的测度是两条后裔序列中不同核苷酸位点的比例：

$$\hat{p} = n_d / n \tag{5-1}$$

这里，n_d和n分别为所检测的两序列间不同核苷酸数和配对总数。在以下的内容中，我们将此估计称为核苷酸间的p距离。

2. 核苷酸替代数的估计　如同氨基酸替代，当序列间亲缘关系较近时，p距离可用来估计每个位点上的核苷酸替代数。然而，当p较大时，因为没有考虑回复突变和平行突变，替代数将被低估。由于核苷酸在序列中只有4种状态，这个问题对核苷酸序列比对氨基酸序列估计更为严重。

估计核苷酸替代数，一般应用核苷酸替代的数学模型。为此，许多学者提出了不同的替代模型，其中一些模型以替代率矩阵的形式列在表5-1中。

表5-1 核苷酸替代模型

(A) Jukes-Cantor模型					(B) Kimura模型				
	A	T	C	G		A	T	C	G
A	--	α	α	α	A	--	β	β	α
T	α	--	α	α	T	β	--	α	αg_G
C	α	α	--	α	C	β	α	--	αg_C
G	α	α	α	--	G	α	β	β	--

(1)Jukes和Cantor方法：这个最简单的核苷酸替代模型由Jukes和Cantor提出。该模型假定任一位点的核苷酸替代都是以相同频率发生的，且每一位点的核苷酸每年以α概率演变为其他3种核苷酸中的一种。因此，一个核苷酸演变为3种其他核苷酸的任何一种的概率为$\gamma = 3\alpha$，γ为每年每个位点的核苷酸替换率。

在这个模型中，我们假设每对核苷酸的替代率相同，所以A、T、C和G的期望频率是0.25。因此，应用公式(5-1)是不需要假定核苷酸频率不随时间变化的。

(2)Kimura两参数法：在实际数据中，转换替代速率常高于颠换速率。Kimura考虑到这种情况，提出一种估计每个位点核苷酸替代数的方法。该模型中，位点转换替代率(α)不同于颠换替代率(2β)。

用Kimura模型，每个核苷酸的平衡频率为0.25。因此，无论核苷酸初始频率为何，均可应用。这一点和Jukes-Cantor模型类似，使得这两个模型较其他模型应用范围更广。

【例5-1】人与猕猴的细胞色素b基因间的核苷酸替代数估计

动物线粒体DNA中的细胞色素b基因是高度保守的，因此常被用于研究亲缘关系较远的动物的进化关系。表5-2列出了人与猕猴的细胞色素b基因的10种不同类型核苷酸对的数目，并分别以密码子第1、2和3位点列出。

表5-2 人和猕猴线粒体细胞色素b基因DNA序列中观察到的10种核苷酸对

密码子的位置	转换		颠换				相同对				总数	
	TC	AG	TA	TG	CA	CG	TT	CC	AA	GG	n_d	n
第1	21	22	5	1	5	4	68	93	100	56	58	375
第2	20	3	6	1	0	2	140	87	71	45	32	375
第3	60	16	6	5	49	2	11	122	102	2	138	375
合计	101	41	17	7	54	8	219	302	273	103	228	1125

表5-3列出了3种不同方法得出的核苷酸替代数估计值$\hat{d}$。对第2密码子来说，4种方法所获得的4种$\hat{d}$值十分接近，$\hat{p}$仅略低于相应的$\hat{d}$值。这表明当$\hat{p}$不大时，不论运用何种方法，同一位点上多重替代的校正实际上并不影响$\hat{d}$值。第1密码子上由4种方法获得的4个估计值$\hat{d}$彼此也相似，虽然它的$\hat{d}$值已接近第2密码子$\hat{d}$值的2倍。然而，在第3密码子上，$\hat{p}$值已充分大，因此多重替代的校正变得不重要。

表5-3 人和猕猴的线粒体细胞色素b基因中第一、第二和第三密码子位置上每位点的替代数估计值

密码子的位置	*p*	Jukes-Cantor	Kimura
第1	15.5 ± 1.9	17.3 ± 2.4	17.8 ± 2.5
第2	8.5 ± 1.4	9.1 ± 1.6	9.2 ± 1.7
第3	32.8 ± 2.5	50.6 ± 4.9	52.3 ± 5.4

（二）氨基酸序列进化分析

1. 氨基酸差异和不同氨基酸的比例　蛋白质或肽链的进化演变研究开始于两个或多个氨基酸序列的比较。这些不同序列分别来自不同的物种。图5-1显示了人、牛、小鼠、大鼠和鸡的血红蛋白α链的氨基酸序列。图中，不同的氨基酸分别用不同的单字母代表。

```
[人]  MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKV
[牛]  MVLSAADKGNVKAAWGKVGGHAAEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGAKV
[小鼠] MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHGKKV
[大鼠] MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVSPGSAQVKAHGKKV
[鸡]  MVLSAADKNNVKGIFTKIAGHAEEYGAETLERMFTTYPPTKTYFPHFDLSHGSAQIKGHGKKV
[人]  ADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASL
[牛]  AAALTKAVEHLDDLPGALSELSDLHAHKLRVDPVNFKLLSHSLLVTLASHLPSDFTPAVHASL
[小鼠] ADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPADFTPAVHASL
[大鼠] ADALAKAADHVEDLPGALSTLSDLHAHKLRVDPVNFKFLSHCLLVTLACHHPGDFTPAMHASL
[鸡]  VAALIEAANHIDDIAGTLSKLSDLHAHKLRVDPVNFKLLGQCFLVVVAIHHPAALTPEVHASL
[人]  KFLASVSTVLTSKYRD
[牛]  KFLANVSTVLTSKYRD
[小鼠] KFLASVSTVLTSKYRD
[大鼠] KFLASVSTVLTSKYRD
[鸡]  KFLCAVGTVLTAKYRD
```

图5-1　五种脊椎动物血红蛋白α链的氨基酸序列

一个简单的测度是两序列间的氨基酸差异数（n_d）。如果所有序列的氨基酸数目相同（n），上述差异数就可用来比较不同序列对间的分歧程度。实际上，当比较很多序列时，氨基酸序列常含有插入或缺失（图5-1）。在这种情况下，计算n_d时一定要删除所有的插入/缺失（间隔）。否则，不同的序列对间相比较时计算出来的n_d是没有意义的。

实际上，不同蛋白质间序列分歧更方便的测度是两个序列间有差异的氨基酸所占的比例。即使n随不同序列而变化，该比例值（p）也可用于比较分歧程度。公式为：

$$\hat{p} = n_d / n \tag{5-2}$$

这一比例值也可称为p距离。假如所有氨基酸位点都以相等概率替代，则n_d遵循二项分布。

表5-4　不同脊椎动物血红蛋白α链中不同氨基酸的数目(上对角线)及不同氨基酸的比例(下对角线)

	人	牛	小鼠	大鼠	鸡
人		16	20	25	42
牛	0.113		19	32	41
小鼠	0.141	0.134		22	41
大鼠	0.176	0.225	0.155		50
鸡	0.296	0.289	0.289	0.352	

注：计算排除了缺失和插入，使用的氨基酸总数为142。

在图5-1所给出的例子中，删除所有间隔后可比较的总氨基酸位点数为140。因此，在此例中n=140。n_d值出现在表5-4对角线上部，可以很容易地计算出$\hat{p}$，列于对角线下部。当所比较的物种亲缘关系很远时(如人和鸡)，$\hat{p}$值较大。这说明随着两个物种的分歧时间增大，氨基酸的替代数也增大，但p并不严格与分歧时间(t)成比例(图5-2)。

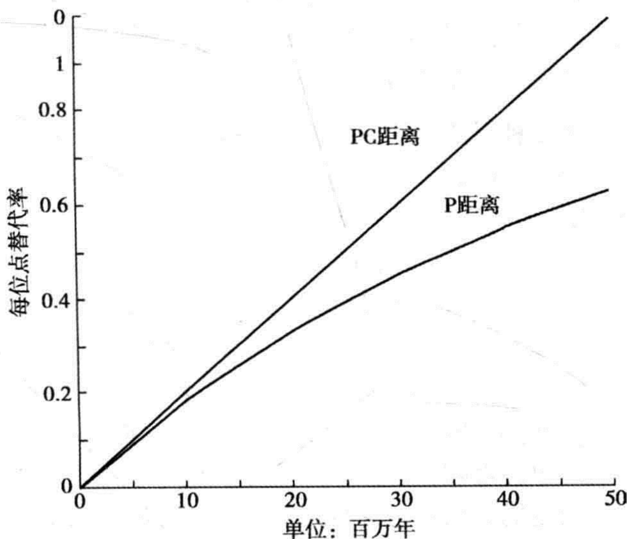


图5-2　p距离和泊松校正(PC)距离随分歧时间(t)变化的关系

2. 泊松校正(Poisson correction，PC)距离　p与t的变化呈现非线性关系，原因之一是当多个氨基酸替代出现在同一位点时，n_d偏离实际氨基酸的替代数将会逐渐增加。运用泊松分布能够更精确估计替代数的方法之一是运用泊松分布的概念。令r为一个特定位点每年的氨基酸替换率(简便起见，假设所有位点的r都相同)，在t年后，每个位点氨基酸替代的平均数为rt。在一个给定位点氨基酸替代数k(k=1,2,3,...)的发生频率遵循泊松分布，即：

$$P(k;t)=e^{-rt}(rt)^k/k! \qquad (5\text{-}3)$$

因此，在某一位点氨基酸不变的概率是$p(0;t)=e^{-rt}$。如果多肽链的氨基酸为n，不变氨基酸的期望值为ne^{-rt}。

实际上，人们并不知道祖先物种的氨基酸序列。因而，只能对已有t年分化的两个同源

序列进化比较来估计氨基酸的替代数。由于一个序列的氨基酸无替代概率为e^{-rt}，因而两个序列同源位点均无替代的概率是：

$$q=(e^{-rt})^2=e^{-2rt} \tag{5-4}$$

此概率可用$1-\hat{p}$来估计，而$q=1-p$。公式中$q=e^{-2rt}$是近似的，因为回复突变和平行突变（在两个不同进化系内出现所导致的同源氨基酸发生同一种突变的情况），并未加以考虑。当然，除非$\hat{p}$相当大（如 > 0.3），上述突变的作用一般可以忽略。

如果应用公式（5-4），则两个序列间每个位点氨基酸替代总数（$d=2rt$）为：

$$d=-\ln(1-p) \tag{5-5}$$

分子进化研究中，常常需要知道氨基酸的替代率（r）。如果从其他生物学信息中已弄清了两个序列间的分化时间t，此速率的估计值为：

$$\hat{r}=\hat{d}/(2t)$$

注意，此处$\hat{d}$被$2t$而不是t所除，因为该速率指一个进化系的速率。

3. 自展法的方差和协方差　可以有若干种方法来估计两个序列间氨基酸替代数。实际上，每个模型都是对真实情况的模拟，仅仅提供了氨基酸的近似替代数。因此，前述的估计距离方差的分析公式也是近似的。用最小二乘法估计多个序列构建的系统树的分支长度时，也需要获得不同序列间的距离方差和协方差的估计值。解决这一问题的一个简便途径是应用自展法（bootstrap）计算多种距离测度的方差和协方差。自展法不要求关于$\hat{d}$值分布的假设，只要求每一个位点是独立进化。

假定有3个是有进化关系的且均含n个氨基酸的序列

$$x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, \dots, x_{1n}$$
$$x_{21}, x_{22}, x_{23}, x_{24}, x_{25}, \dots, x_{2n}$$
$$x_{31}, x_{32}, x_{33}, x_{34}, x_{35}, \dots, x_{3n}$$

这里，x_{ij}表示第i个序列第j个位点上的氨基酸。对序列1、2，序列1（、）与3以及序列2、3分别计算$\hat{q}$值，即$\hat{q}_{12}$、$\hat{q}_{13}$和$\hat{q}_{23}$。把$\hat{q}_{ij}$代入公式，便获得序列i和j的PC距离（$\hat{d}_{ij}$）。

在自展法计算方差和协方差时，具有n个氨基酸的3个序列的随机样本是从原始数据集中产生的。随机样本以伪随机数从原始的数据集中按列有放回随机抽取，形成自展重复抽样数据集。一旦获得了随机样本，便能对3对序列的每一对计算出距离的估计值。如此重复B次，便能产生B个距离值$\hat{d}$。以$\hat{d}_b$表示第b次自展重复抽样的$\hat{d}$值，然后可用式（5-6）计算自展方差：

$$V_B\left(\hat{d}\right)=\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{d}_b-\bar{d}\right)^2 \tag{5-6}$$

这里，$\bar{d}$是所有重复抽样$\hat{d}_b$的平均值。一般来说，计算$V_B(\hat{d})$可做约1000次重复抽样（B=1000）。

自展法通常基于一个假设，即所有位点都是独立进化。在位点总数低时，这一假设是不成立。但如果位点总数很大（n > 100），如本例中，此假设可以成立，因为以不同速率替代的大多数位点在每次自展样本上都会出现。

自展法的一个优点是，在没有数学公式可用时，也能算出方差和协方差，而且能比近似的数学公式提供更好的估计。它能方便地以同样的标准统计公式对任何距离测度计算出方差和协方差。但是，当原始样本太小且存在偏倚时，这种偏倚不能被自展法消除。在这种情况下，解析法将得到比自展法更准确的方差和协方差。

【例5-2】由解析法和自展法获得的PC距离标准误

表5-5列出了由解析式和自展法算出的PC距离（$\hat{d}$）的标准误，自展法重复了1000次。它们均基于图5-1的血红蛋白α链数据。表5-5列出了上述数据集的$\hat{d}$值。显然，由上述两种方法所获得的标准误基本是一致的。对*p*和Γ距离，用上述两种方法也可以获得几乎相等的标准误。因此，用自展法估计进化距离的标准误是合适的。

表5-5　解析法估算的PC距离的标准误（下对角阵）
及自展法估算的PC距离的标准误（上对角阵）

	人	马	牛	袋鼠	蝾螈	鲤鱼
人		0.031	0.031	0.039	0.078	0.083
马	0.031		0.030	0.043	0.083	0.081
牛	0.031	0.031		0.038	0.080	0.079
袋鼠	0.040	0.043	0.039		0.081	0.084
蝾螈	0.074	0.080	0.076	0.080		0.090
鲤鱼	0.082	0.081	0.079	0.086	0.089	

4. Γ距离　以上所介绍的进化距离都有一个假定，即所有核苷酸位点的替代速率相同。事实上，速率可因位点不同而变化。在蛋白质编码基因中，密码子的第1、第2和第3个位置上的替代率是不同的。蛋白质活性中心的氨基酸功能制约也对氨基酸位点间的速率差异有重要影响。在RNA编码基因上也观察到速率差异现象，主要是由于RNA功能限制及二级结构的影响。不同位点替代速率的统计分析指出，速率变异近似地遵循Γ分布。

鉴于上述原因，许多学者致力于发展适用于核苷酸替代的Γ距离。一般而言，Γ距离比非Γ距离更符合实际，但前者比后者方差更大。有鉴于此，除非所使用的核苷酸数目非常大，否则Γ距离不一定对构建系统树有更优的结果。

二、系统发育树重建方法

在研究从病毒到人类的各种生物的进化历史中，DNA或蛋白质序列的系统发育分析已经成为一个重要的工具。由于不同的基因或DNA片段的进化速率存在较大的差异，我们可以通过这些基因或DNA片段来估计几乎所有水平上的有机体间的进化关系。系统发育分析对于阐明多基因家族的进化关系，以及理解在分子水平上的适应性进化过程也是十分重要的。

（一）系统发育树的种类

1. 有根树和无根树　基因或生物体的系统发育关系常常用有根或无根的树形结构来表

示，即有根树和无根树。树的分支样式称为拓扑结构。对一定规模的分类群（任何分类学单位：属、种、群体和DNA序列等），可能的有根树和无根树的拓扑结构数目很大。如果一个类群数为m的有根二叉树，其可能的拓扑结构数为：

$$1 \cdot 3 \cdot 5 \cdots\cdots (2m-3)=[(2m-3)!]/[2^{m-2}(m-2)!], (m \geq 2)$$

若m=10，则有34 459 425有根二叉树。无根树可能的拓扑结构的计算来用m-1替换公式中的m即可，即m=10时，结果为2 027 025种。在大多数情况下，大部分可能的拓扑结构可以通过明显不可能的进化关系或其他信息排除。

2. 基因树和物种树　进化学家常常对代表一个物种或群体进化历史的系统发育树感兴趣，这种树称为物种树或种群树。然而，当一个系统发育树由来自各个物种的一个同源基因构建时，得到的树将不完全等同于物种树。当某一座位出现等位基因多态性时，从不同物种取样的基因分离的时间将比物种分歧时间长。根据基因构建的树的分支结构也可能不同于物种树，我们称这种树为基因树。同样需要注意的是，如果检测的氨基酸或核苷酸数目较少，重建的基因树和物种树的分支式样也可能不同。因此，可以通过检测大量的氨基酸或核苷酸来避免这种错误。

当所研究的基因属于一个多基因家族时，有可能出现问题。因为构建一个不同物种的系统发育树，我们应当使用直系同源而不是旁系同源，因为只有直系同源才代表物种形成事件。然而，事实上，要区分直系同源基因和旁系同源基因是很难的。

3. 期望树与现实树　在推断系统发育的理论中，常常假设所研究的DNA或蛋白质序列非常长（理论上无限长），从中获得的大量核苷酸或氨基酸均是随机取样。一个用无限长的序列或每一分支的替代数的期望值构建的树称为期望树，建立在实际替代数基础上的树称为现实树，由所观察到的序列数据构建的树称为重建树。期望树、现实树和重建树通常是不同的。大多数构建树的方法的目的是重建现实树，这一类方法包括邻接法、最大简约法和最大似然法等。

当选择构建树的DNA序列不同，重建树的拓扑结构和分支长度也将不同，因此，评价物种树或种群树时，应尽量使用多基因。

4. 拓扑距离　两个不同的树之间的拓扑距离通常可以用序列分割的方法来测量。对于无根二叉树，这个距离是有差异内部分支数的两倍。如果两个8序列的树具有相同的拓扑结构，则dT=0，若所有内部分支均产生不同的分割，则 dT=10。然而，如果比较的两个树具有多歧点，则上述规则不起作用，这种情况下，我们可以使用Rzhetsky和Nei的普遍性公式计算：

$$d_T=2[\min(q_1,q_2)-p]+|q_1-q_2| \quad (5\text{-}7)$$

这里，q_1和q_2分别为树1和树2的内分支树，p是使两树产生相同序列的分割树。当包含多歧点时，q_1和q_2可能不同；但对于二叉树，q_1和q_2一般是相同的。

（二）基于距离法构建系统发生树

构建系统发生树通常使用的方法分为3大类：①距离法；②简约法；③似然法。

构建树的方法一般包括两个过程：拓扑结构的判断和一个既定的拓扑结构分支长度的估计。当拓扑结构已知时，估计分支长度可以用多种统计学方法，如最小二乘法和最大似然法等，问题在于如何判断或重建一个拓扑结构。

系统发育重建的方法具有很大的争议，曾经从事通过形态学特征来研究系统发育的研究者倾向于使用假设条件较少的简约法；从事分子生物学工作的研究者倾向于使用分析法；

数学家和统计学家试图建立各种复杂的数学模型,而较少地考虑实际应用。

距离方法:距离方法涉及两个步骤:计算物种对之间的遗传距离以及从距离矩阵重建一棵系统发育树。下面我们介绍两种不需要分子钟假设的方法:最小二乘法(least-squares,LS)和邻接法(neighbor-joining,NJ)。

(1)最小二乘法(图5-3):最小二乘法将成对距离矩阵作为给定数据,通过匹配那些尽可能近的距离来估计一棵树上的分支长度,即对给定的和预测的距离差的平方和最小化。预测距离是沿连接两个物种的通路的分支长度总和计算的。距离差的平方和的最小值则是树与数据(距离)相似测度,它可用作树的分值。

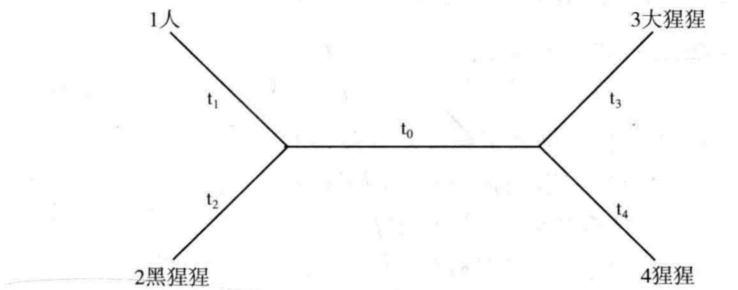


图5-3 估计枝长的最小二乘标准的示意图

设物种i和j之间的距离为d_{ij},树上物种i到j间通路的枝长和为$\hat{d}_{ij}$。LS方法对所有独立的i和j对求距离差的平方$\left(d_{ij}-\hat{d}_{ij}\right)^2$的最小值,使得这棵树与距离之间的拟合尽可能地近。例如:对Brown等的线粒体数据在k80模型下计算成对距离(见表5-6)作为观测数据。现在,考虑树人,黑猩猩,大猩猩,猩猩及它们的5个枝长t_0、t_1、t_2、t_3、t_4。

表5-6 线粒体DNA序列的成对距离

	1. 人	2. 黑猩猩	3. 大猩猩	4. 猩猩
1. 人				
2. 黑猩猩	0.0965			
3. 大猩猩	0.1140	0.1180		
4. 猩猩	0.1849	0.2009	0.1947	

在这棵树上,人与黑猩猩之间的预测距离是t_1+t_2,人与大猩猩之间的预测距离是$t_1+t_2+t_3$,依此类推。则距离差的平方和为:

$$S=\sum_{i<j}\left(d_{ij}-\hat{d}_{ij}\right)^2$$
$$=\left(d_{12}-\hat{d}_{12}\right)^2+\left(d_{13}-\hat{d}_{13}\right)^2+\left(d_{14}-\hat{d}_{14}\right)^2+$$
$$\left(d_{23}-\hat{d}_{23}\right)^2+\left(d_{24}-\hat{d}_{24}\right)^2+\left(d_{34}-\hat{d}_{34}\right)^2$$

表5-7 K80模型（Kimura，1980）下的最小二乘法

树	t_0	t_1	t_2	t_3	t_4	S_j
τ:((*H*, *C*), *G*, *O*)	0.008840	0.043266	0.05328	0.058908	0.135795	0.000035
τ:((*H*, *C*), *C*, *O*)	0.000000	0.046212	0.05623	0.061854	0.138742	0.000140
τ:((*H*, *C*), *C*, *O*)	同上					
τ:((*H*, *G*), *C*, *O*)	同上					

在表5-7中S是5个未知枝长t_0、t_1、t_2、t_3、t_4的函数。最小化S的枝长值为LS估计：$\hat{t}_0$=0.008840，$\hat{t}_1$=0.043266，$\hat{t}_2$=0.053280，$\hat{t}_3$=0.058908，$\hat{t}_4$=0.135795，对应的树分值为S=0.00003547。对其他几棵树，可以进行类似的计算。的确，其他几棵二元树都趋向于星状树，内分支长估计值为0。具有最小S的树称为LS树，它是真实系统发育关系的LS估计。

用最小二乘标准确定的树采用同样的标准估计分支长，计算一个散点图中与$y=a+bx$配合的直线。如果对枝长没有什么约束，就有解析解，可以通过解线性方程获得。非约束方法可以是树重建的一种良好的方法，但是对枝长没有明确定义。一些模拟研究建议约束枝长为非负值，将改善树重建效果，大多数计算机程序在现实LS方法时不采用约束。值得注意的是，当所估计出的枝长为负值时，它们多数时候其实是接近于0。

（2）邻接法：对树进行比较（特别是距离法中）所用的一个标准是以树的枝长总和来度量进化总量，枝长总和最小的树称为最小进化树（minimum evolution tree）。

邻接法是基于最小进化标准的一种聚类算法。由于它计算快，又能产生合理的树，因而得以广泛应用。它从一个星状树开始，然后加入两个节点，选择能达到树长减少最大的一对。随后，产生一个新节点来替代两个加入的节点将矩阵的维数减少了一次。重复这一过程，直到完全解出这棵树，该算法的每一步都要更新树的枝长以及树长。

（三）基于字母特征构建进化树

最大简约法：在采用等位频率来重建人类种群间的关系时，研究者建议进化树的合理估计为进化总数的最小值，这种方法在应用于离散数据时被称为简约法，而最小进化法在今天被看做是对重复突变进行修正后枝长总数最小化的方法。

在一个位点上性状变化的最小数目常常被称作性状长度（character length）或位点长度（site length）。对序列上的所有位点而言，性状长度之和是对整个序列所需要变化的最小数目，称为树长（tree length）、树分值（tree score）或简约分值（parsimony score）。具有最小树分值的树是真实树的估计，称为最大简约树。多棵树是等价最佳树的情况经常见到，尤其是序列非常相似时。

假设在某个特定位点，4个物种的数据是AAGG，且考虑图5-4给出的两棵树所需的最小变化数目。我们通过将性状状态标注到灭绝的祖先状态节点来计算这个数目。见图5-4。

对第一棵树，可以通过标注A和G到两个节点来做到这一点，内枝只需要一次变化（A-G）。对第二棵树，我们可以将AA（已显示）或GG（未显示）标注到两个内节点，任何一种情况下，最少都需要两次变化。注意，某位点上被标注为祖先状态的一组性状状态被称为祖先重建（ancestral reconstruction）。对于具有（n-2）个内节点的n物种的二元树而言，在每个

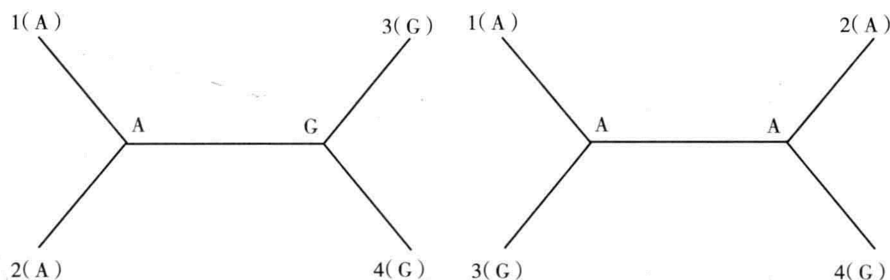


图5-4 最大简约法建树示意图

位点重建的总数为4(n-2)(核苷酸)或20(n-2)(氨基酸)。达到变化最小数目的重建称为最简约重建(most parsimonious reconstruction)。因此,对第一棵树,只有一个单一的最简约重建,而对第二棵树,两个重建是等价最简约。

一些位点对树的判别并无贡献,因而是没有信息的。例如一个恒定位点,即所有物种在该位点具有相同的核苷酸,对任何树都不影响。类似地,单变位点——即两个观察的性状中有一个只出现一次(例如,TTTC或AAGA)——对每棵树只需要一次变化,因而也不是信息位点。一个性状为AAATAACAAG(对10个物种)的位点也是非信息的,因为对任意树只要对所有祖先节点标注A都需要3次变化。对一个简约信息位点(parsimony-informative site)而言,至少要有两个状态被观测到,每一个至少两次。注意,信息位点和非信息位点的概念仅仅只用于简约法。而在距离法或似然法中,所有位点(包括不变位点)都影响计算,应当被包括在内。

我们常常将所有物种在某个位点上观察到的性状状态看做是位点构型(site configuration)或位点模式(site pattern)。这意味着对4个物种而言只有3种位点式样是有信息的,它们是*xxyy*, *xyxy*和xyyx,这里*x*和*y*是任意两个不同状态。很明显,这3种位点式样分别"支持"3棵树,分别是*T*1 :((1,2),3,4); *T*2 :((1,3),2,4)和*T*3 :((1,4),2,3)。设具有这些位点式样的位点数分别是*n*1, *n*2和*n*3,如果*n*1, *n*2或*n*3是3个中最大的,则*T*1, *T*2和*T*3是最简约树。

(四) 用于系统发育重建的距离测度

1. 当每个位点的核苷酸替代数目的Jukes-Cantor估计值小于0.05时,应当使用p距离或Jukes-Cantor距离,而不管是否存在转换/颠换,不管替代速率是否因核苷酸位点而异。

2. 当0.05<d<1,且检验的核苷酸较多时,用Juker-Cantor距离,除非转换/颠换比较高(R>5)。但此比率较高且检测的核苷酸数目很多时,要使用Kimura距离。

3. 对于很多序列来说,d>1时构建的系统树会因为某些原因而不可靠(如存在对位排列错误),因此,建议尽量避免使用这些数据。可以淘汰进化很快的那部分基因区域(如去除免疫球蛋白的超变区基因),仅使用进化速度慢的区域。

4. 当距离很大而n很小时,用来估计每个核苷酸位点替代数据的很多距离方法不能使用,在这种情况下,p距离可以获得相对可靠的拓扑结构。

5. 当一个系统树是通过一个基因的编码区构建时,同义与非同义替换之间的差别就很重要,可以用dS来构树。

6. 普遍地,如果两种距离测度对于同一数据获得相同的距离值(或极为相近)时,应该

使用简单的测度，因为它的方差较小。

三、分子钟假说

(一) 概述

分子钟（molecular clock）假说认为DNA或蛋白质序列的进化速率随时间或进化谱系保持恒定。在20世纪60年代初期，人们就观察到不同物种中蛋白质序列的差异，如血红蛋白、细胞色素C及血纤肽中大致与物种分歧时间成正比。通过这些观察，提出了分子进化钟的概念。

首先需要澄清几点。第一，分子钟应当被看做是氨基酸或核苷酸突变的随机性所导致的随机钟。它不像普通钟表以固定时间间隔跳动，而是以一个随机间隔跳动。第二，不同蛋白质间或蛋白质的不同区域间进化速率的差异很大，因而分子钟假说允许不同蛋白质间进化速率不同，或者说每个蛋白质有其自身固有的分子钟，以不同的速率跳动。第三，速率恒定性未必对所有物种适用，很有可能只存在于某一类群中。例如，我们可以说就某个特定基因而言，分子钟假说在灵长类中成立。

在分子进化的中性学说（neutral theory of molecular evolution）提出之时，分子进化的“似钟特性”被认为“可能是该学说最有力的证据”。中性学说强调相对适应度接近于零的中性或近中性突变的随机固定。分子进化的速率则等于中性突变率，而与环境变化或种群大小等因素无关。如果突变率相似而蛋白质功能在同一类群中保持不变，以至于中性突变比例相同，那么根据中性学说的预测，进化速率将是恒定的。蛋白质间的速率差异则被解释为由于不同蛋白质具有不同的功能限制，因而中性突变的比例不同。

近年来，考古学数据被用来校定分子钟，即将序列间的距离转换成绝对地质时间和置换率。病毒基因分析涉及类似的情况，其进化非常迅速，以至于数年之内就可以观测到变化。人们可以用病毒被隔离的时间来校正分子钟，并使用与这里讨论基本相同的方法来估计分歧时间。

(二) 相对速率检验

最简单的分子钟假设检验是采用第三个物种C（外类群）来检验两个物种A和B是否以相同的速率进化。这一检验称为相对速率检验（relative-rate test），其实几乎所有的分子钟检验比较的都是相对速率而不是绝对速率。如果分子钟假说为真，那么从祖先节点O到物种A和B的距离应当相等：$d_{OA}=d_{OB}$和$a=b$。同理，人们可以得出$d_{AC}=d_{BC}$。

(三) 内部分枝检验

1. 正态偏离（Z）检验　如前所述，推断树的可靠性是通过检验其每个内部分枝的可靠性来完成的。这个检验（内部分支检验）适用于由距离法构建的树。考虑5序列树，在5序列的情况下，有15种可能的无根二分歧树，每个树由5个外部分支和2个内部分支组成。假设拓扑结构A是正确的，而其他的都是不正确的，则表明正确拓扑结构的所有分支长度估计的期望值是0或者正值，而不正确拓扑结构中至少有一个内部分支长度为负值，且该分支产生了序列间的一个不正确分区。只要使用无偏距离估计而分支长度用LS方法估计，则对于任何

数目的序列构造的树进行检验似乎都是正确的。因此，如果一个树的某个内部分支估计值被确定为负值，该树的拓扑结构很可能就是错误的。

上述的零假设检验能相当方便地应用于由距离法（特别是由NJ或者ME方法）获得的树的分析上，因为只有正确树的所有内部分支才可能是正值的。但在MP和ML树中，不管拓扑结构如何，所有内部分支都为正值，因此，就很难建立出一种检验零假设的分析方法。然而，使用自展法可以检验零假设。

2. 自展内部分支检验　另一种用于距离树内部分支检验的是自展内部分支检验。这种方法是检验一个给定树的每个内部分支的可靠性。与自展检验法相似，从原始序列中随机抽样形成与原始数据数目相同的核苷酸（或者氨基酸），再用从原始序列数据获得的树拓扑结构来计算所有分支长度，并对同一种拓扑结构重复数百次。一个内部分支的长度估计b将随着重复次数变化而不同，且可能为负值。我们可以计算b的平均数以及标准误，并进行Z检验。

该检验结果通常与上述分析方法获得的结果非常相似。但是该方法优于解析法，即无需分别计算每个替代模型b的标准误；所有替代模型的标准误可用同样的方法计算。因此计算时间不会随序列数增加而迅速增加。这个方法比解析法更易运用。然而，当核苷酸或者氨基酸数目小时，该方法可能会给出Pc的有偏估计，这是因为如果原始样本有偏差，则此偏差在重复抽样时不能被除去。在这种情况下，解析法要好得多。

第二节

核苷酸和蛋白质的适应性进化

Section 2 Adaptive Evolutions of Nucleotide and Protein

基因和基因组的适应性进化最终决定形态、行为和生理上的适应，以及物种分歧和进化创新(evolutionary innovation)。因此，在分子进化研究中，分子适应是一个令人振奋的课题。尽管自然选择在形成形态和行为进化方面似乎普遍存在，但它在基因和基因组进化中所起的作用尚存在争议。分子进化的中性学说认为，种内和种间大多数可见差异不是由自然选择，而是由适合度很小的随机突变的固定决定的。40年来人们发展了一系列中性检验方法，本节介绍正选择和负选择的基本概念以及分子进化的主要理论，还将简要介绍几种群体遗传学中发展起来的常用的中性检验方法。另外引入应用范围比较广的dN/dS检验，并且详细介绍了其计算方法。

一、中性与近中性理论 ❯❯❯

在群体遗传学中，一个新突变基因a与野生型显性基因A的相对适合度由选择系数s来度量。设基因型AA，Aa和aa的相对适合度分别为1，1+s和1+2s，则s<0，=0及>0分别对应负选择(negative selection)或净化选择(purify selection)、中性进化和正选择(positive selection)。新突变基因的频率各世代高低不同，既受自然选择又受随机漂变的影响。究竟是随机漂变还是自然选择决定了突变的命运取决于Ns(N为有效群体的大小)。若$|Ns|>>1$，则自然选择决定基因命运；若$|Ns|$接近于0，则随机漂变的作用非常重要，而且该突变为中性或近中性。

按照中性理论，我们今天观察到的遗传变异——无论是种内多态性还是种间分歧，均不取决于自然选择所驱动的有利突变的固定，而是取决于那些事实上没有适合效应(即中性的)突变的随机固定。下面是该理论的一些观点和预测。

(1)大多数突变是有害的，会被净化选择所清除。

(2)核苷酸置换率等于中性突变率(即总突变率乘以中性突变所占比例)。如果物种间中性突变率恒定(或者日历时间或者世代时间)，则置换率也是恒定的。这个预测为分子钟假说提供了解释。

(3)功能较重要的基因或基因区域进化较慢。在具有较重要作用或处于较强功能约束下的一个基因中，中性突变比例较小，使得核苷酸置换率较低。现在，功能重要性和置换率之间的负相关在分子进化中是一个普遍现象。例如，替代置换率几乎总是比沉默置换率低；

密码子第3位比第1和第2位进化更快；具有相似化学性质的氨基酸比不相似的氨基酸更容易相互替代。如果自然选择在分子水平上驱动进化过程。那么可想而知，功能重要的基因的进化速率比功能不重要的基因要高。

（4）种内多态性和种间分歧是中性进化同一过程的两个阶段。

（5）形态特征（包括生理、行为等）的进化的确是自然选择所驱动的。中性学说关注的是分子水平上的进化。

围绕中性理论的争论已产生很多的群体遗传理论和分析工具。下面将讨论其中几种。

二、微观适应性进化的检验方法

以下几个是典型的统计学研究适应性进化的方法，已经形成了稳定的软件。根据输入数据的不同可以检验相应基因的选择强度。

1. Tajima的*D*检验　在随机交配的群体中，一个中性基因上保持的遗传变异量由$\theta = 4N\mu$决定，这里N为（有效）群体大小，μ为每一代的突变率。从每个位点的角度定义θ，它也是从群体中随机抽取的每条序列的期望位点杂合度。例如，在人类非编码DNA中，$\hat{\theta} \sim 0.0005$，意味着两条随机的人类序列间大约0.05%的位点不同。群体数据一般很少有变异，所以通常采用无限位点模型，假定每个突变都发生在DNA序列的不同位点上，且无须校正多重命中。注意，群体规模大和突变率高都会导致群体中保持更高的遗传变异。

两种从群体中随机抽取DNA序列的简单方法可以用来估计θ。第一种是包含*n*条序列的样本中的多态性位点数S，期望值$E(S) = L\theta a_n$，这里的L为序列中的位点数，$a_n = \sum_{i=1}^{n-1} 1/i$，故$\theta$可由$\hat{\theta}_S = S/(La)$估计。第二种方法是对*n*条序列所有成对比较的核苷酸差异的平均比例值的期望为θ，将θ作为一个估计值，则记作$\hat{\theta}_\pi$。这两种θ的估计在中性突变模型下均无偏，即假定无选择、无重组、无群体分化或大小变化，以及突变和漂变之间平衡。然而，如果模型的假设不成立，则不同因素对$\hat{\theta}_S$和$\hat{\theta}_\pi$有不同影响。例如，若轻微有害突变在群体中保持较低频率能显著增加S和$\hat{\theta}_S$值，但对$\hat{\theta}_\pi$几乎没有影响。θ的两个估计量可以为了解造成严格中性模型失效的因素和机制提供信息。因此，Tajima构建了以下的检验统计量：

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_S}{SE\left(\hat{\theta}_\pi - \hat{\theta}_S\right)} \tag{5-8}$$

这里，*SE*为标准误差。

在无效中性模型下，*D*的均值为0，方差为1。Tajima建议采用标准正态分布和β分布来确定*D*是否显著不同于0。

Tajima的*D*检验的统计显著性可能与几种不同的解释相容，而且难以区分它们。正如前面所讨论的，一个负*D*值表明存在净化选择或群体中分离的轻微有害突变。然而，负*D*值也可能是由群体扩张造成的。在一个扩张群体中，可能分离出许多新的突变，且它们在数据中

以单元(singletion)形式出现,即其他所有序列在此位点上都相同,只有一条序列不同。单元增加了分离位点的个数并导致D值为负。类似地,D值为正可解释为平衡选择将突变维持在居中频率。然而,一个收缩的群体也能够导致D值为正。

2. Fu和Li的D检验与Fay和Wu的H检验　在n条序列的一个样本中,一个多态位点上突变核苷酸的频率为$r=1, 2, \cdots, n-1$。样本中观察到的突变的这种分布成为位点频谱(site-frequency spectrum)。通常,采用亲缘关系很近的外类群来推断祖先的和衍生的核苷酸状态。例如,若在一个$n=5$的样本中观察到的核苷酸为AACCC,而外类群中为A(假定的祖先状态),则$r=3$。Fu设r为突变规模。如果祖先状态未知,则不可能区分突变规模是r还是$n-r$,使得那些突变被划为同一类,位点频谱则被认为是折叠的,折叠构象提供的信息远少于非折叠构象,因而,采用外类群来推断祖先状态应当增加检验效力,但缺点是该检验可能会受到祖先重建中误差的影响。

Fu和Li区分了内部突变和外部突变,即分别在系谱树内枝或外枝上发生的突变。设这两类突变的个数分别为η_I和η_E,注意η_E为单突变的个数,他们构建了以下的统计量:

$$D=\frac{\eta_I-(a_n-1)\eta_E}{SE\left(\eta_I-(a_n-1)\eta_E\right)} \tag{5-9}$$

这里,$a_n=\sum_{i=1}^{n-1}1/i$,SE为标准误差。与Tajima D检验相类似,该统计量也是作为中性模型下θ的两个估计值间的差异来构建的。Fu和Li认为群体中分离的有害突变倾向于近期产生,位于树的外枝,且对η_E起作用;而内枝上的突变多为中性,且影响η_I。

3. McDonald-Kreitman检验和选择强度估计　中性学说认为种内多样性(多态性)和种间分歧是同一进化过程的两个阶段,即两者都是由中性选择突变的随机漂变所致。因而,如果同义和非同义突变都是中性的,则种内同义和非同义多态性的比例应与种间同义和非同义差异的比例相同。

近缘物种蛋白质编码基因中的可变位点可依位点是否具有多态性或固定差异,以及该差异是同义还是非同义的,划分为一个2×2列表中的4类(表5-8)。假设我们从物种1中抽取5条序列,从物种2中抽取4条序列,若某位点在物种1中数据为AAAAA,在物种2中为GGGG,则该差异被称为固定差异。若某位点在物种1中的数据AGAGA,而在物种2中为AAAA,则该位点被称为多态性位点。注意,无限位点模型无需对隐藏变化进行校正。如果数目不多,则中性无效假设等价于列表的行和列之间独立并可被χ^2分布或Fisher精确检验验证。MeDonald和Kreitman测定了果蝇3个亚群的乙醇脱氢酶基因(*Adh*)序列,获得了表5-8中列出的数据。P值小于0.006,说明与中性期望有显著偏差。种间替代突变远多于种内替代突变。McDonald和Kreitman将此模式认作驱动种间差异的正选择证据。

表5-8　果蝇Adh基因中存在沉默突变、置换突变以及多态性位点个数(数据来自McDonald and Kreitman,1991)

变化类型	固定差异	多态性
置换(非同义)	7	2
沉默(同义)	17	42

为了弄清这个解释后面的推论，假定同义突变是中性的，考虑选择对物种分歧之后出现的非同义突变的影响。人们预期有利替代突变会很快固定下来并成为种间的固定差异。因而，若固定的替代突变过剩（如同在*Adh*中观察到的），则表明存在正选择。

人们在哺乳动物线粒体基因中已观察到过剩的替代多态性，表明净化选择下存在轻微有害替代突变。有害突变被净化选择清除，而且不会在种间比较中看见，但在种内还是会分离。

三、宏观适应性进化的检验方法

蛋白质编码序列区分为同义置换和非同义置换，对理解自然选择的作用来说，这比内含子或非编码序列优越得多。若将同义置换率作为基准点，我们可以推断自然选择在非同义置换固定过程中是推动还是阻碍作用。非同义/同义置换率的比率（$\omega = d_N / d_S$）可以在蛋白质水平度量选择压力。如果选择对适合度没有影响，则非同义突变将以与同义突变相同的速率被固定，使得dN=dS及ω=1。如果非同义突变是有害的，则净化选择将降低其固定速率，使得dN<dS及ω<1。如果非同义突变受到达尔文选择的青睐，则其被固定的速率将高于同义突变，致使dN>dS及ω>1。因此，非同义突变率显著高于同义突变率即为蛋白质适应性进化的证据。

然而，可以预料一个功能蛋白上的大多数位点在大部分进化时间都是受约束的。即使发生正选择，也只能影响几个位点，且只有偶尔发生。因此，这种成对平均方法很少检测到正选择。近期研究着重检测影响系统发育关系中特定谱系或蛋白质中单个位点的正选择。

对编码蛋白质的DNA序列，同义和非同义置换被定义为平均每个同义位点上的同义置换数（*ds*或*Ks*）以及平均每个非同义位点上的非同义置换数（d_N或K_A）。

本节主要使用记数法计算，计数方法类似于JC69等核苷酸置换模型下的距离计算，有3个步骤：①对同义和非同义位点计数；②对同义和非同义差异计数；③计算差异比例并校正多重命中（multiple hit）。将位点和差异都计数后，就可以区分同义和非同义这两种类型间的差异了。

1. 位点计数　每个密码子都有3个核苷酸位点，分成同义和非同义两类。以密码子TTT（Phe）为例，由于3个密码子位置上每个核苷酸都可以转变为另外3种核苷酸，该密码子就有9个直接邻居：TTC（Phe），TTA（Leu），TTG（Leu），TCT（Ser），TAT（Tyr），TGT（Cys），CTT（Leu），ATT（Ile）和GTT（Val）。其中，密码子TTC和密码子TTT编码同一个氨基酸。因此，对密码子TTT而言，就有3×1/9=1/3个同义位点，3×8/9=8/3个非同义位点（表5-9）。在计数过程中，不计入变为终止密码子的突变。我们将该方法用于序列1中的所有密码子，并将计数结果相加以获得全序列中同义和非同义位点的总数。然后，对序列2重复该过程并计算两条序列间的平均位点数目，分别计为*S*和*N*，有$S+N=3 \times L_c$，这里L_c为序列中的密码子的数目。

表5-9　密码子TTT（Phe）中的位点计数

目标密码子	突变类型	置换率（*K*=1）	置换率（*K*=2）
TTC（Phe）	同义	1	2
TTA（Leu）	非同义	1	1

续表

目标密码子	突变类型	置换率(K=1)	置换率(K=2)
TTG(Leu)	非同义	1	1
TCT(Ser)	非同义	1	2
TAT(Tyr)	非同义	1	1
TGT(Cys)	非同义	1	1
CTT(Leu)	非同义	1	2
ATT(Ile)	非同义	1	1
GTT(Val)	非同义	1	1
总和		9	12
同一位点数		1/3	1/2
非同义位点数		8/3	5/2

注：K 为转换/颠换置换率比率。

2. 变异计数　第二步是对两条序列间的同义和非同义变异进行计数。换言之，在两条序列间所观测的差异可按同义和非同义划分。我们再按密码子逐一处理。很明显，如果两个所比较的密码子相同（如TTT对TTT），则同义和非同义变异数目为0；如果两个所比较的密码子间仅在一个位置上存在差异（TTC对TTA），就很容易发现这种单一的变异是同义的还是非同义的。然而，如果两个比较的密码子间在2~3个位置上都存在差异（如CCT对CAG或GTC对ACT），则有4~6条进化途径能使一个密码子变成另一个密码子。多条途径中可能涉及同义和非同义差异数不同。大部分计数方法对不同途径赋予同等权重。

例如，密码子CCT和CAG间存在两条途径（见表5-10）。第一条途径要通过中间密码子CAT转换，涉及两个非同义变异；而第二条途径通过中间密码子CCG转换，涉及一个同义变异和一个非同义变异。如果我们对这两条途径赋予相同权重，则两个密码子间有0.5个同义变异和1.5个非同义变异。如果同义突变率高于非同义突变率，如同几乎所有基因中表现的一样，第二条途径应该比第一条途径的可能性更大，预先不知道d_N/d_S比率和序列分歧度，就很难对不同途径赋予合适的权重。不过，计算机模拟结果表明加权对估计值的影响很小，尤其是当序列的分歧度并不是很大时。

表5-10　密码子CCT和CAG间的两条途径

途　径	差异	
	同义	非同义
CCT(Pro)↔CAT(His)↔CAG(Gln)	0	2
CCT(Pro)↔CCG(Pro)↔CAG(Gln)	1	1
平均	0.5	1.5

计数沿着序列密码子逐一进行，将差异数相加得到两条序列间总的同义和非同义差异数，分别记为S_d和N_d。

3. 多重命中校正 现在，我们有：

$$p_s=S_d/S$$
$$p_N=N_d/N \tag{5-10}$$

分别是同义和非同义位点上的差异比例，它们等同于针对核苷酸的JC69模型下的差异比例。因此，我们套用JC69中对多重命中的校正。

$$\begin{aligned} d_s &= -\frac{3}{4}\log\left(1-\frac{4}{3}p_s\right) \\ d_N &= -\frac{3}{4}\log\left(1-\frac{4}{3}p_N\right) \end{aligned} \tag{5-11}$$

当我们只关注同义位点和差异时，每个核苷酸并不存在3个其他核苷酸来突变的情况。实际上，对多重击中校正的作用很少，至少在序列分歧度不高时如此，故校正公式带来的偏差也就不是非常重要了。

4. rbcL基因应用实例 我们应用上述方法来估计黄瓜和烟草中叶绿体蛋白1,2-二磷酸核酮糖羧化酶/加氧酶大亚基（*rbcL*）基因间的d_s和d_N。黄瓜（Cucumissativus）*rbcL*基因的Genbank序列号为NC _ 007144，烟草（Nicotiana tabacum）为Z00044。在黄瓜和烟草基因中分别有476个和477个密码子，对位排列后的序列则有481个密码子。我们删除了任意一个物种对位排列时出现的间隔密码子，这样序列中就剩下472个密码子。

表5-11列举了数据的一些基本统计值，它们是对3个密码子位置分别进行分析后获得的。碱基组成不等，第三个密码子富含A/T。3个密码子位置的转换/颠换置换频率的比率估计值大小依次为$\hat{K}_3>\hat{K}_1>\hat{K}_2$。序列距离的估计值也是同样的顺序$\hat{d}_3>\hat{d}_1>\hat{d}_2$。这类模式在蛋白编码基因中很常见，反映了遗传编码结构以及基本上所有氨基酸都处于选择压力之下，同义置换率高于非同义置换率。当对密码子逐一进行检测时，两个物种间有345个密码子是一致的，115个密码子在一个位置上有差异，其中95个是同义的，20个是非同义的。10个密码子在两个位置上有差异，2个密码子在3个位置上均不相同。

表5-11 黄瓜和烟草rbcL基因的基本统计量

位置	位点	π_T	π_C	π_A	π_G	$\hat{\kappa}$	$\hat{d}$
1	472	0.179	0.196	0.239	0.386	2.202	0.057
2	472	0.270	0.226	0.299	0.206	2.063	0.026
3	472	0.423	0.145	0.293	0.139	6.901	0.282
总计	1416	0.291	0.189	0.277	0.243	3.973	0.108

随后，1416个核苷酸位点被分为S=343.5个同义位点以及N=1072.5个非同义位点。在两条序列间观察到141个差异，这些差异分为S_d=103.0个同义变异和N_d=38.0个非同义差异。因此，在同义和非同义位点上的差异比例分别为$p_s=S_d/S=0.300$和$p_N=N_d/N=0.035$。使用JC69校正后得到d_S=0.383和d_N=0.036，其比值$\hat{\omega}=d_N/d_S=0.095$。根据这一估计，该蛋白处于强烈

的选择压力之下，在群体中发生一个非同义突变的概率只有同义突变的9.5%。

四、适应性进化基因

基于ω比率检验获得的大多数正选择基因可分为以下3类。第一类包括针对病毒、细菌、真菌和寄生虫攻击的防御机制或免疫作用中的宿主基因，以及与破坏宿主防御机制有关的病毒或病原基因。例如，前者包括主要组织相容性复合体、淋巴细胞蛋白CD54、植物中与识别病原有关的R基因及哺乳动物中反转录病毒抑制剂TRIM5α；后者包括病毒表面或包膜蛋白、疟原虫细胞膜表面抗原以及由植物天敌（如细菌、真菌、卵菌、线虫和昆虫）产生的多糖。可以想见，病原基因由于受到正选择进化出不被宿主防御机制识别的新类型，同时宿主也必须适应并识别出病原，这就激发了一场进化"军备竞赛"，驱动新的替代突变在宿主和病原中固定。蛇或蝎子毒液中的毒素用于捕获猎物，也处于类似选择压力下，因而进化速率很快。

第二类主要包括与生殖有关的蛋白质或信息素。一批研究已检测到有关精-卵识别蛋白质及雄性或雌性生殖其他方面的快速进化。这些基因上的自然选择也可能加速或导致新物种形成。

第三类正选择基因与上述两类有所重叠，包括基因复制后获得新功能的基因。基因复制是基因、基因组和遗传系统进化的初级驱动力，被认为在新基因功能进化中起引领作用。复制基因的命运由能否为机体带来选择优势所决定，多数复制基因被清除或因有害突变失去功能而退化为假基因。由于亲代基因需要不同功能，有时新拷贝会在适应进化驱动下获得新功能。已检测到许多基因在基因复制后经历加速蛋白质进化，其中包括灵长类DAZ基因家族、灵长类绒毛促性腺蛋白。群体遗传检验也表明正选择在复制核基因早期进化动态中的重要作用。

还有很多其他基因也被检测处于正选择之下，尽管它们不如那些参与到进化军备竞赛中的基因（如宿主-病原拮抗作用及生殖）那么多。这也许是基于ω比率的检验方法的局限性所致，即可能错过一次性的适应性进化。在这种进化中，一个有利突变出现并迅速在群体中扩散开来，接踵而至的就是净化选择。若要检测到更多正选择，也许需要改进能检测影响某个谱系上少数位点的插曲式或局部的进化方法。

统计检验不能证明基因是否真正经历适应性进化。具有信服力的例子也许要建立在实验验证和功能检验上，两者在观察到的核酸变化与蛋白质折叠以及表型变化（如催化化学反应的效率不同）之间建立直接联系。

第三节 分子进化与生物信息学

Section 3 Molecular Evolution and Bioinformatics

一、基因组进化概述

基因组学（Genomics）是一门只有十多年历史的新兴学科，发展极为迅速，并产生了许多分支学科。随着研究的不断深入，它已从结构基因组学（structural genomics）进入到功能基因组学（functional genomics）。利用基因组学研究的方法和成果来研究生物进化，也就是进化基因组学（evolutionary genomics）所要研究的问题，越来越受到进化生物学研究者的关注。

目前，尽管进化基因组学还没有正式列在基因组学的议事日程上，但也已经有了不少相关的研究，比较基因组学（comparative genomics）就是其中之一。对不同生物基因组结构的异同及其特点进行比较，除了在功能基因组学的研究上很有意义外，还有可能在一定程度上了解基因组的进化，特别是基因组的结构特征与生物复杂性的关系。例如，通过比较，发现基因组中蛋白质和功能RNA基因的密度与生物的复杂程度有一定的负相关。在细菌基因组中，基因的平均密度是1个基因/1kb; 在酵母中，是1个基因/2kb; 而线虫是1个基因/5kb; 果蝇是1个基因/13kb; 到人类则是1个基因/40kb。这种密度的变化显然是与基因组进化中调控元件和“非基因序列”的扩增有关。

比较基因组学的研究还表明，基因和基因组是由并非很多的基本结构单位（构件）构成的，而这些构件在进化中被反复使用（重组）。以形成新的基因和基因组，这就像为数不多的化学元素可以组成无数的化学物质（分子）那样。新的化学分子是通过已有元素或分子之间的化学反应产生的，所以，基因组的进化有可能以化学反应作为其动态模型，即新基因组的产生是通过已有基因或基因组的重组、重排、重新建立新的关系而达成。要充分认识这种类比的意义，就必须开展进化基因组学的研究。

基因组的进化与基因组的三维结构之间显然也有很重要的关系。人与黑猩猩DNA序列的相似程度达99%，两者的差异很可能是在其基因组的三维结构（包括三维调控关系）上。因此，进化基因组学必将深入进行这方面的研究。

为了了解基因组及其发展变化的本质，当然还要研究与生命起源有关的最原始的基因和基因组的起源，以及其后的进化模式与过程，这样，我们就有可能在分子水平上认识生物进化的分段途径。总之，进化基因组学将是基因组学中最触及事物本质的一个分支。

二、病毒基因组分析

对生物的分类应该体现其系统演化。对病毒来说，它的生命是相对脆弱的，很难达到像古细菌、细菌和真核生物那样综合全面的程度。病毒也受突变和自然选择的影响，并且病毒基因组的进化速度远远超过其他细胞的基因组。有很多证据证明，早在一万年前病毒就已经存在，这些证据包括人类的骨骼残骸，历史记录和遗物。然而，远古病毒的DNA或RNA还没有被找到。

RNA病毒基因组的RNA聚合酶一般缺乏校正能力。这导致基因组的突变率比DNA基因组高100万~1000万倍。对于DNA病毒，其突变率一般比宿主细胞高10~1000倍。除了高突变率，许多病毒的复制速度也是极其惊人的。单个细胞能产生10 000个脊髓灰质炎病毒颗粒，而一个被艾滋病病毒感染的个体一天能产生10亿个病毒颗粒。许多病毒的基因组由相对独立的多个片段组成。这些片段能够在病毒复制过程中随机重组，从而在子代病毒中产生大量不相同的子类。流感病毒几乎每年都能引起大范围的疾病流行就是这个原理的体现。病毒经常处于强大的选择压力下，如宿主的免疫反应或抗病毒药物作用。因此，艾滋病病毒快速的突变和复制确保某些病毒株通过突变产生对抗病毒药物的抗性，而且会经受环境的选择而存活下来。

病毒经过漫长的进化历程已经能够侵入系统发生树中所有物种: 古细菌、细菌和真核生物。植物病毒（番茄丛矮病毒）、动物病毒（如SV40病毒，鼻病毒和脊髓灰质炎病毒）以及噬菌体（如噬菌体ΦX174）的衣壳蛋白中都有“β-折叠桶”或“果冻卷”折叠结构。除非发生了显著的趋同进化，否则这种现象一般说明这些病毒是同源的。感染植物和动物的反转录病毒具有双链RNA基因组以及封装它的特殊衣壳体。有一类噬菌体（Φ6）也具有这种特征，也说明了感染不同物种的病毒之间具有同源性。在对这些病毒基因组以及蛋白质的分析中并没有发现序列相似性，再次凸显了病毒基因组高速进化的特点。病毒基因组的高度多样性使我们无法根据其序列数据绘制出涵盖所有病毒的全面完整的系统发生树，这反映了病毒基因组形成历程中复杂的分子进化事件。

三、原核生物基因组比较

(一) 与人类疾病相关的细菌分类

细菌和真核生物已经相互“交战”几百万年了。细菌为了繁殖需要占据人体这个营养丰富的环境。典型的细菌“殖民地”包括皮肤、呼吸道、消化道（口腔、大肠）、尿道和生殖系统等。据估计每个人身上的细菌数目超过自身的细胞数目。大多数情况下，这些细菌对人类是无害的。然而，有些细菌在一定条件下能够导致感染，甚至带来灾难性的后果。最近一些年，由于广泛使用抗生素导致了细菌抗药性的增强，因此急需找到细菌的毒性因子，然后找到相应的接种疫苗。对这个问题的一个解决办法就是比较细菌的致病株和非致病株。

(二) 原核生物基因组比较数据库

NCBI提供了一个非常有效的基因组比较工具，并且使用起来非常容易。从基因组查询

页面上，选择果蝇（Drosophila melanogaster）就得知到图5-5所示的页面。选择TaxPlot，就能够将两个基因组和一个参考基因组（如caenorhabditis elegans和saccharomyces cerevisiae）进行比较。在这个图上，每一个点都代表参考基因组中的一个蛋白质。x坐标和y坐标显示了被比较蛋白质组中每个蛋白质最佳匹配的BLAST分值。如果蛋白质都在图的对角线上，表明它们在参考蛋白和输入蛋白中的分值相同（或者几乎相同）。然而，也有值得注意的异常值，代表了两种生物不同表型的重要基因。这些点是可以点击的（图中带圆圈的数据点）。TaxPlot还能根据COG分类系统规则在图上标注颜色。

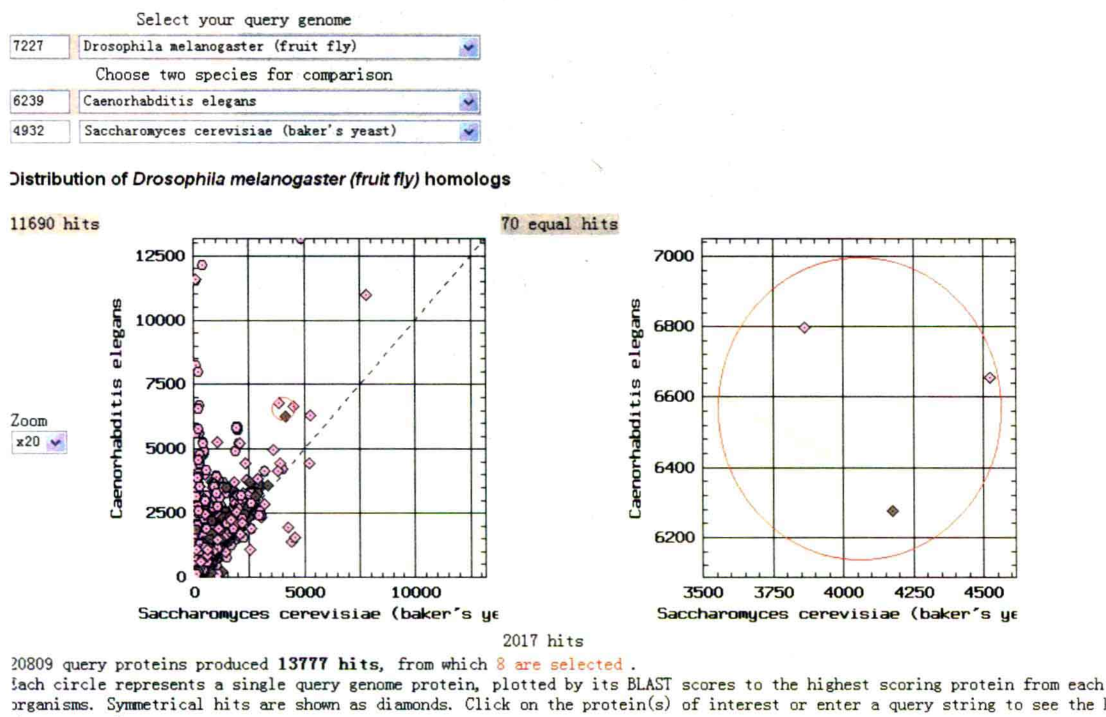


图5-5　Taxplot界面示意图

在整个微生物基因组的比对中最大的挑战就是来用动态程序，比对上百万的碱基对所需要的大量时间。然而对于基因组比对来说，这些工具还比较初级。MUMmer软件包提供了一个对微生物基因组进行快速准确比对的方法。最近，经过对算法改进后，也能够对真核生物序列进行比对。

MUMmer将两条序列作为输入。这个算法找到了所有的长于一个设定的最小长度值*k*并且很好匹配的子序列。根据定义，这些匹配序列是最小的，因为如果将它们向任意方向延长一点就会导致不匹配。

MUMmer的输出结果由点阵图组成（图5-6），该结果以最小比对长度150bp为序，显示了两个基因组序列的比对结果。结果包括如下内容：SNPs；比单个SNP更加分散的序列区域；大的插入片段（例如，经过转座、序列逆转和水平基因转移）；散在重复片段（例如，一个基因组中的复制）；片段串联重复（拷贝数）。

大肠埃希菌K12和大肠埃希菌O157：H7（在受污染的食品中有这个菌株，会导致如出血性结肠炎之类的疾病）。在大约45亿年前发生分枝。测序并比较两个基因组，发现大肠埃希

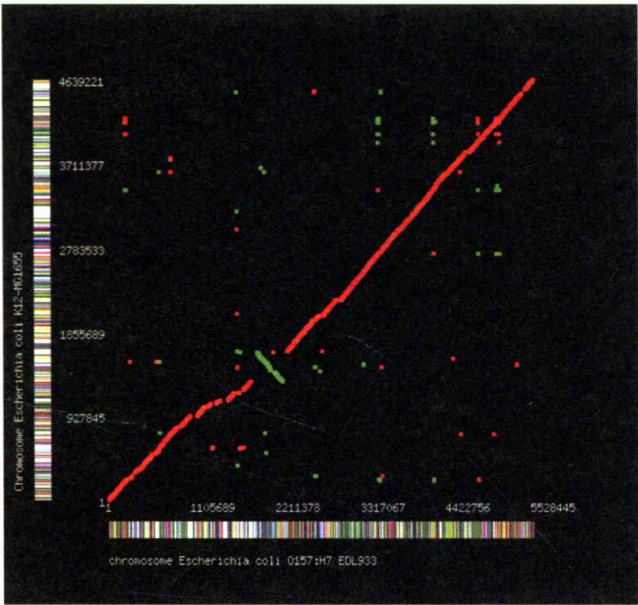


图5-6 MUMmer输出结果

菌O157：H17大约比大肠埃希菌K12长了859 000个碱基对。这两个细菌有大约4.1Mb的共同基因组骨架，大肠埃希菌O157：H7有另外1.4Mb的序列（大部分通过水平基因转移得到）。MUMmer的输出结果对于找出两个基因组中的共同区域和反向重复区域非常有用。

四、蛋白质互作网络进化

近年来，随着鉴别蛋白质互作关系的高通量实验技术（如酵母双杂交，免疫共沉淀，基于质谱的串联亲和纯化等）以及生物信息学方法在预测蛋白互作领域的发展与应用，越来越多的蛋白质互作数据涌现出来，为进化研究提供了新的视角。

对蛋白质互作网络的进化分析可分为五个层面：蛋白质个体、蛋白质互作对（protein interaction pair）、模体（motif）、网络模块（network module）以及整个网络。即按照包含蛋白质的数目将网络进化问题分层：第一层是仅包含一个蛋白的蛋白质个体；第二层为包含两个蛋白的蛋白互作对；网络模体一般包含3~5个蛋白质，为第三层；网络模块作为第四层，相对于之前的三层包含的蛋白数目更多，且可能由模体组成；第五层则是整个网络的进化分析，探究网络的发生发展过程。

（一）网络中的蛋白质个体进化

蛋白质互作网络对蛋白质个体进化性质的影响，即蛋白质互作是否会减慢蛋白质进化速率，是在蛋白质个体层面上研究网络进化的主要问题。

由于研究者选择的研究对象多数为酵母，尽管所选的互作数据不同，采用的进化速率评估方法、寻找直系同源蛋白的方法及所统计分析方法等不尽相同，但从现有的研究成果可以得出如下结论：蛋白连接度同其进化速率之间可能存在较弱的负相关关系。因为影响蛋白质进化速率的因素很多，除了与网络拓扑性质相关的蛋白连接度（由互作数目定义），蛋白中

心性（由介数定义）外，还有可能与蛋白表达水平，蛋白必要性，蛋白质功能及其参与的生物学过程，蛋白质丰度，密码子适应指数等有关，并且这些因素之间存在错综复杂的依赖关系。

（二）网络中的蛋白互作对进化

互作的两个蛋白质在进化上是否趋向具有相似的性质？在分子水平上是否趋向共进化？这是网络中蛋白互作对进化研究要回答的问题。

多年来，研究者开发了许多预测蛋白质互作的方法，如比较基因组学方法、利用系统发育树相似性进行预测的方法、利用基因表达水平相关性进行预测的方法和同源预测方法等，这些方法多是基于相互作用蛋白共进化的思想。这些预测算法的成功，从另一个角度为互作蛋白具有共进化的现象提供有力证据。目前学术界普遍认同的观点是：互作的蛋白质倾向于具有更相似的进化速率，且网络中的蛋白互作对在表达水平等层次上也可能存在微弱的共进化现象。对于这一观点的解释主要有两种，一种假设为，共进化是施加在互作的蛋白对上相似进化压力的结果。相似的进化压力可能来源于作用在这两个互作蛋白对上的相似调控机制，如协同转录和调控等。这种假设不仅适用于解释发生直接物理互作蛋白对间的共进化，对共享一个生物学关系的一组蛋白质的共进化现象也同样适用。另一种假设为，共进化直接与互作蛋白的共适应相关。即当蛋白序列上直接或者间接通过影响蛋白质折叠而参与互作的位点发生有害突变时，与其互作的蛋白通过发生互补的改变来维持两蛋白的互作关系，进而保持功能。综合两种假设，即两种共进化推动力可能是在不同程度，不同水平和不同情况下发挥各自的作用。

（三）网络中的模体进化

网络模体是指复杂网络中在不同位置重复出现的特定的相互连接模式，在数量上显著地高于随机期望，一般含有3~5个节点。对于网络模体进化的研究主要集中在探讨模体是否对其成员蛋白进化具有约束作用。研究表明，模体成员蛋白要比非模体成员蛋白在进化上更具有保守性。在不同拓扑结构模体中，成员蛋白的保守性不同，可能的原因是不同的模体模式所承受的进化约束显著不同。

（四）网络中的模块进化

蛋白质互作网络具有层次模块化特性。功能模块的最显著特点是其往往表现出可能在功能和拓扑上互相联系，在蛋白互作网络中主要以蛋白质复合物的形式存在。目前的研究成果表明，网络的模块化对蛋白质进化可能有约束作用，成员蛋白之间在进化速率，表达水平等方面表现出共进化特性。类似蛋白质互作预测领域，许多功能模块预测算法（如比较基因组学方法）都是基于模块成员蛋白共进化的思想，其成功也反过来支持了功能模块成员蛋白的共进化特点。

（五）网络的整体进化

研究蛋白质互作网络整体进化的最主要问题是蛋白质互作网络的起源。随之而来的问题是蛋白质互作网络具有的无尺度（scale-free）分布，小世界（small world）性质和模块化结构等是如何起源和进化的？这些特性的存在是生物体长期进化过程中自然选择的结果，还

是存在内在约束机制使其发生成为不可避免的趋势?

多年来,学者们先后提出了多个无尺度和小世界网络的进化模型。目前应用最为广泛的是优先连接模型和复制-分歧模型。优先连接模型描述网络的生长是通过不断向网络中添加新的节点来实现的,而新添加的节点倾向于优先与原有网络中度高的节点连接。这一模型揭示的问题是蛋白质年龄与连接度之间存在的强烈而显著的关系,即蛋白质起源越早,其连接度越高。并且当控制表达水平后,这种关系并没有被显著地削弱。在复制-分歧模型中,网络中的初始蛋白质被随机选择并复制,且伴随该蛋白质参与的所有互作。随后,基因突变导致副本和原蛋白逐渐发生分歧,表现为它们参与的互作发生改变。从生物信息学的角度,则可以理解为基因组层面上的改变在网络拓扑结构变化上的体现。有研究表明,酵母中至少有40%的蛋白质互作来源于复制事件。而对于蛋白质复合物的起源和进化研究显示,有相当一部分复合物是通过逐步的部分复制而进化来的,并且被复制的复合物仍然保持原复合物的核心功能,但具有不同的绑定特异性和规则,具体见图5-7。

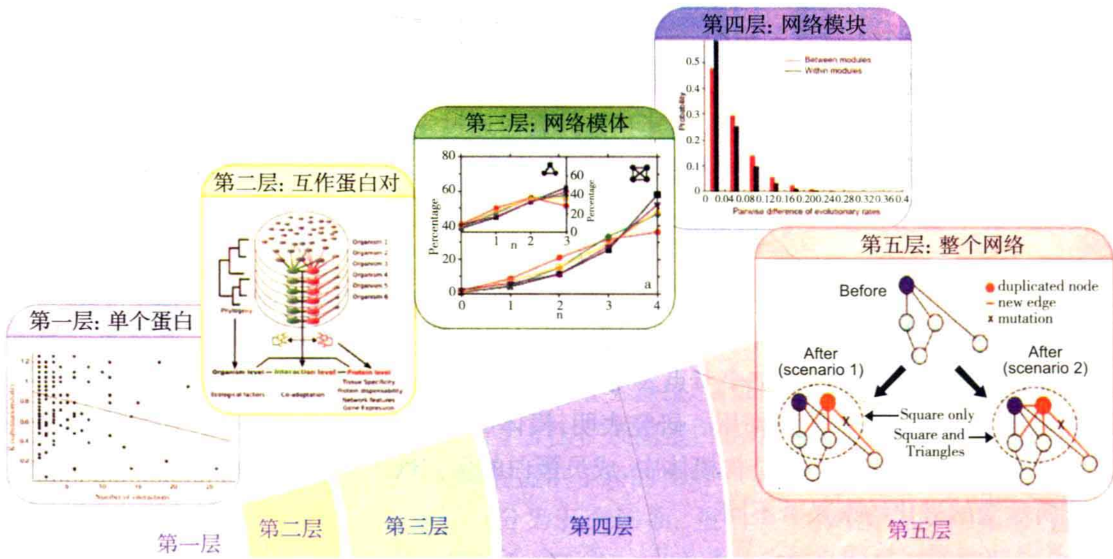


图5-7 蛋白质互作网络进化图

第一层表示网络中的蛋白质个体进化,表明蛋白连接度同其进化速率之间存在较弱的负相关关系。第二层表示网络中的蛋白互作对进化,揭示出互作的蛋白质倾向于具有更相似的进化速率可能由多种因素导致。第三层表示网络中的模体进化,模体成员蛋白更具有保守性。第四层表示网络中的模块进化,成员蛋白之间在进化速率表现出共进化特性。第五层表示网络的整体进化中的复制-分歧模型

五、代谢网络进化分析

各种高通量技术和代谢通路数据库的发展使得分析代谢网络进化(metabolic network evolution)成为可能。一般说,生物网络具有稳健性和进化性的一个主要原因归功于其模块化组织。模块定义为一组连接非常紧密的基因或酶的集合,功能相对独立,而模块与模块之间的连接较为稀疏。从仅有几个基因的简单网络能够利用计算机模拟的手段构建出具有几

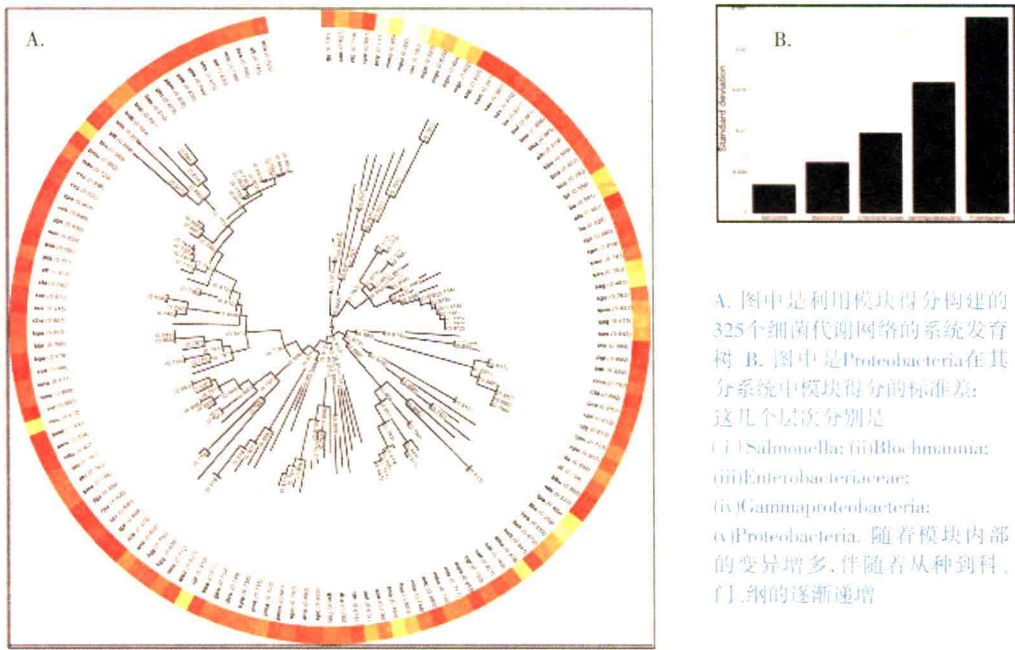
百个节点上千条边的大网络。另外，有些研究通过比较多个物种的拓扑结构对代谢网络的进化机制进行探讨，发现不同代谢通路的拓扑特征提供不同的系统发育信息。

(一) 代谢网络模块性的进化分析

一个生物网络中的模块包含很多元素（例如蛋白质或反应），这个模块形成了一个结构上的子系统，并且有其独特的功能。在代谢网络中，存在很多小的，高连接度的模块，这些模块又分层组合成为大的单元。对于模块的进化，目前主要有两个假设：一是模块倾向于正选择，因为已经限定好的模块能维持细胞的功能，通过模块的进化变化能够提升其可进化性；二是尽管模块不能直接通过选择进化，但模块之间在进化上存在一致性，还能通过其他可以被选择的性质，例如由水平基因转移引起的基因聚类的加速，多效性的最小化，和对新环境的适应性等。

由于生物之间的遗传相关，其代谢网络也存在着一定的相似性，所以系统发育相近的生物代谢网络模块也应该是相近的。伴随模块内变异逐渐增多，物种之间的差异也就越大，相反亦然。如果对不同物种代谢模块统计相应得分，就可以根据这个得分构建生物代谢系统发育树。但对模块的变异量化研究存在一定难度，如何计算每种生物代谢网络的得分是研究关键。

Anat Kreimer等人成功解决了这个问题，他们根据模块的特性，使用Newman的算法计算代谢网络中模块的得分，根据每个物种计算得到的代谢模块分数建立距离矩阵，形成了如下图（图5-8）所示的系统发育树。



A. 图中是利用模块得分构建的325个细菌代谢网络的系统发育树 B. 图中是Proteobacteria在其分系统中模块得分的标准差：这几个层次分别是（ⅰ）Salmonella；(ii)Blochmannia；(iii)Enterobacteriaceae；(iv)Gammaproteobacteria；(v)Proteobacteria. 随着模块内部的变异增多，伴随着从种到科、门、纲的逐渐递增

图5-8 利用代谢网络模块得分建立其系统发育树

(二) 代谢与环境互作的进化分析

代谢网络一般是在一定的生化环境下行使功能，同时通过吸收和分泌各种有机和无机

的化合物来与环境发生互作。例如在网络内部新陈代谢流动性的分布或生命体的增长率都是通过这种作用来完成。

和环境的这种相互作用在一定程度能够在代谢网络的结构进化上反映，所以这些代谢网络不应只是单单推断代谢功能，还应当能够观察到物种和环境互作进化的现象。在分析代谢网络的拓扑结构时，有一类化合物是通过外源获得，这类化合物定义为“种子集合”。如果一个物种的环境能够决定其代谢反应，那么这些“种子集合”就是代谢网络与外界环境之间一个很好的代理（图5-9）。

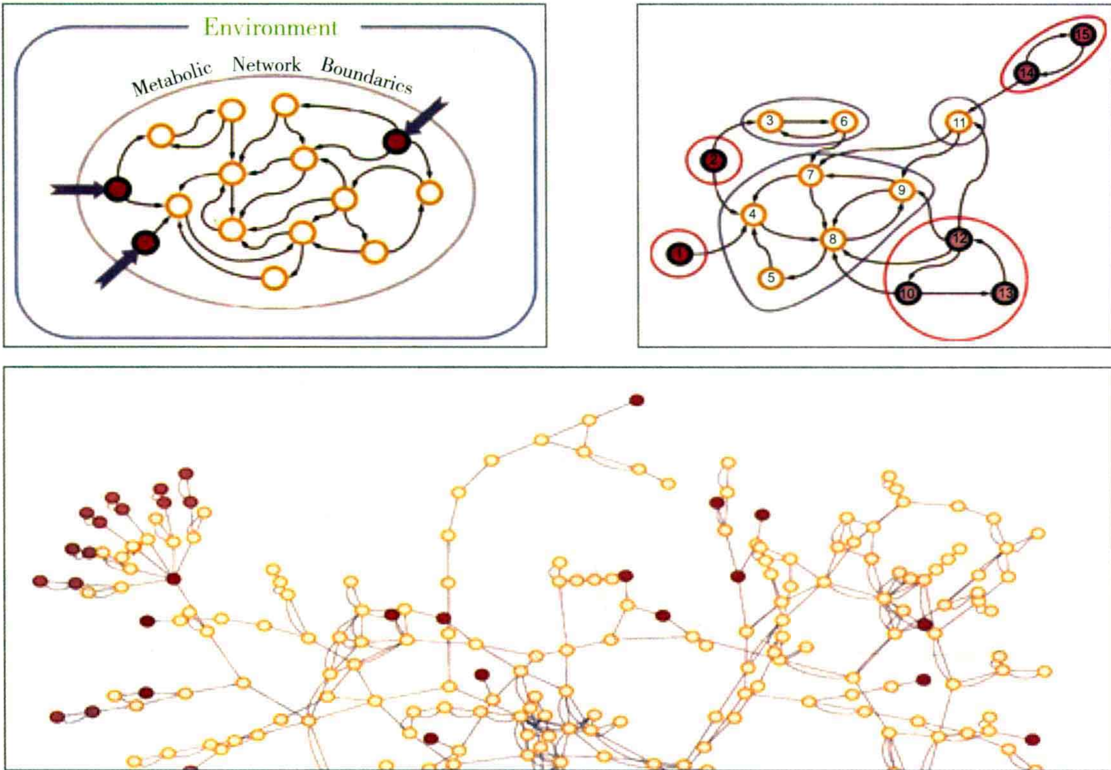


图5-9　代谢与环境互作的进化分析示意图

在代谢网络中鉴定种子复合：A.代谢网络与环境相互作用的示意图，种子是用红色标记；B.代谢网络中种子获得过程。网络首先用kosaraju的强连通组分（SCC）的方法分解，子网中的源组分就是要找的种子。图中的源组分是用红色表示的，节点颜色的饱和程度代表种子的置信程度；C.Buchnera代谢网络图，红色为种子复合物

每种生物的代谢网络种子集合是不同的，根据集合中的基因在这种生物是否存在可以构造进化的距离矩阵。因为在进化过程会有新的化合物以种子或者非种子的身份加入到代谢网络中，如果是以种子的身份被整合到代谢网络中，这个种子存在的状态可能不会太长，要么从代谢网络中被拿掉，要么快速变为非种子化合物。

第四节

应用实例：SARS流行病的系统发生分析

Section 4 Reconstructing the Origin and the Diffusion of the SARS Epidemic

2003年2月28日，暴发一场大规模的流行系疾病，经确认，命名为急性呼吸系统综合征（Severe Acute Respiratory Syndromes，SARS）。同年3月15日，WHO发布全球警告，称SARS为"世界范围的健康威胁"。他们警告可能的地点包括加拿大、印度尼西亚、菲律宾、新加坡、泰国和越南。

流行病的起源：尽管SARS的起源和原因还不知道，但应该离我们知道的时间不远，我们通过分析多个SARS基因组就可以知道这个疾病是怎样发生和它的起源以及如何在许多国家扩散的。在2003年3月的第3周，美国、加拿大、德国及中国香港分别独立的从SARS患者身上分离出新的冠状病毒（SARS-CoV）。

通过分析大量的完整病毒基因组数据集，可以回答我们很多重要的问题。这里，我们也将提出一些工具来回答这些问题中的一部分。是怎样一种病毒导致了这样一场流行病？这种病毒的原始宿主是什么？跨越物种障碍的时间和地点？是怎样一个关键突变让这种转换成为可能？

为了回答这些问题，我们首先要了解一些系统发生分析关键算法，这些在前面章节中已经提到过，然后把这些算法应用于2003年获得的SARS数据（所有这些数据都可从Genbank获得）。

1. SARS基因组　SARS-CoV基因组是在2003年4月由加拿大团队获得的，29 751bp的单链RNA序列。我们可以通过GenBank获得这个数据（查询编号为AY274119.3）。在图5-10中提供了该病毒的基因图谱。其GC含量大概是41%，是已经公布的冠状病毒基因组GC含量范围之内的。并且由一个典型的冠状病毒结构，按照一定的顺序排列5个或者6个基因。

2. SARS流行发生重构　在SARS流行病发生的时候，有关其起源和本质等许多关键的问题都可以通过基因组序列分析来获得。在2003年早期多个团体就已经获得和发布了SARS的序列，并以此作为基础作为探寻流行病起源和扩散，现在我们可以用GenBank中的许多病毒序列来研究这次流行病。我们选取了13条已知获取时间和地点的序列，然后展示如何用这些序列来挖掘这次流行病的信息。

鉴定宿主：SARS病毒在早期被认为是冠状病毒，和已知其他冠状病毒有相同序列的基因。然而，它又是完全不同于其他已知人类冠状病毒，因此，很可能是从其他动物中起源的。我们使用多种动物冠状病毒的蛋白质构建了邻接树，其中包括了在果子狸中发现的冠状病

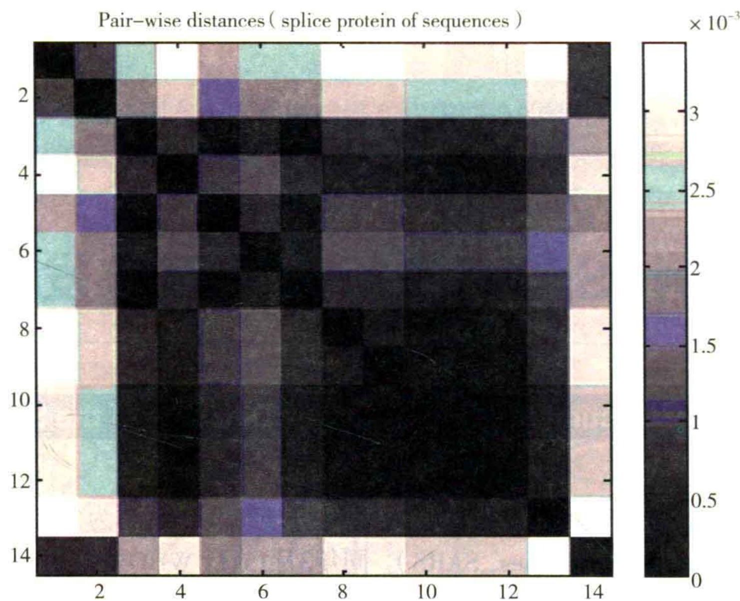


图5-10 SARS病毒两两比对遗传距离

毒。SARS看起来和果子狸冠状病毒最相近，和人类其他冠状病毒都比较远。

使用表5-12的13个基因组，我们用邻接法构建了系统发生树，这种疾病并不是通过鸟类携带的，而是起源于果子狸，然后在人类中传播。这个距离矩阵是通过Jukes-Cantor模型计算的距离矩阵并且用核苷酸序列做全局比对进行校正作为遗传距离。

表5-12 SARS病毒发生时间及地点调查表

Table	Name, location, and sampling date of SARS virus isolates used in our case study		
Name of isolate	Acc.number	Date	Location
GZ01	AY278489	DEC-12-2002	Guangzhou (Guangdong)
ZS-A	AY394997	DEC-22-2002	Zhongshan (Guangdong)
ZS-C	AY395004	JAN-04-2003	Zhongshan (Guangdong)
GZ-B	AY394978	JAN-24-2003	Guangzhou (Guangdong)
HZS-2A	AY394983	JAN-31-2003	Guangzhou Hospital
GZ-50	AY304495	FEB-18-2002	Guangzhou (Guangdong)
CUHK-W1	AY278554	FEB-21-2003	Hong Kong
Urbani	AY278741	FEB-22-2003	Hanoi
Tor 2	AY274119	FEB-27-2003	Toronto
Sin2500	AY283794	MAR-01-2003	Singapore
TW1	AY291451	MAR-08-2003	Taiwan
CUHK-AG01	AY345986	MAR-19-2003	Hong Kong
Palm civet	AY627048		Palm civet

从这棵树上，我们能够了解这次流行病的整个过程。如果把果子狸作为外类群，我们可以看到所有早期的病例都是发生在广东省，并且Hotel Metropole冠状病毒几乎和它们中的一条序列是完全一致。

因为我们已经知道每个测序的SARS病毒收集的时间，这样就能观察到经过若干时间突变的过程。方便起见，我们使用了spike蛋白质对应的开放读码框。相对于从果子狸获得的序列，我们看到其遗传距离随着时间在粗略按线性模式逐渐提高（x轴表示时间，原点代表2003年1月1日）。如果我们在这些数据中插入最小二乘法的拟合曲线，就可以估计这次流行病起源的大概时间。任何一个在零点附近日期都可能是开始的时间，估计在2002年9月16日到2003年1月1日之间。这种方法是比较粗糙的，而且其中很多假设我们还都没有证实，但仍然给我们一个很可能的时间点，最早的病例报道可以追溯到2002年的下半年。

（张绍军）

参考文献

1. Borenstein E. , Kupiec M. , Feldman M. , et al. , *Large-scale reconstruction and phylogenetic analysis of metabolic environments.* Proc Natl Acad Sci U S A, 2008. 105(38): 14482–14487.
2. Chen K, Rajewsky N. *The evolution of gene regulation by transcription factors and microRNAs.* Nat Rev Genet, 2007. 8(2): 93–103.
3. Cristianini N. *Introduction to Computational Genomics A Case Studies Approach.* 2006 : Cambridge university press.
4. Fraser, H. B. , Hirsh A. E. , Steinmetz L. M. , et al. , *Evolutionary rate in the protein interaction network.* Science, 2002. 296(5568): 750–752.
5. Juan D, Pazos F. Valencia A. *Co-evolution and co-adaptation in protein networks.* FEBS Lett, 2008. 582(8): 1225–1230.
6. Kreimer A. , Borenstein E. , Gophna U. , et al. , *The evolution of modularity in bacterial metabolic networks.* Proc Natl Acad Sci U S A, 2008. 105(19): 6976–6981.
7. Nei M. *Molecular Evolution and Phyolgenetics.* 2000 : Oxford university press.
8. Pevsner J. *Bioinformatics and Functional Genomics 2nd Edition.* 2009 : Wiley-Blackwell.
9. Wuchty S. , Oltvai Z. N, , Barabasi A. L. , *Evolutionary conservation of motif constituents in the yeast protein interaction network.* Nat Genet, 2003. 35(2): 176–179.
10. Yang Z. *Computational molecular evolution* Oxford Series in Ecology and Evolution, ed. P. H. H. R. M. May, British: Oxford university press. 2006.

第六章

CHAPTER 6

新一代测序数据分析

NEXT-GENERATION SEQUENCING DATA ANALYSIS

DNA测序技术已广泛应用于生物学研究的各个领域，很多生物学问题都可以借助高通量DNA测序技术予以解决。这几年，大规模平行测序平台（massively parallel DNA sequencing platform）已经发展为主要的测序技术，这项测序技术的出现不仅令DNA测序费用降到了很低，还让基因组测序这项以前专属于大型测序中心所拥有的“特权”能够被众多研究人员分享。同时新一代DNA测序技术有助于人们更全面、更深入地分析基因组、转录组及蛋白质之间交互作用组的各项数据。今后，各种测序将成为一项广泛使用的常规实验手段，这有望给生物学和生物医学研究领域带来革命性的变革。

第一节

全基因组测序与重测序

Section 1 Whole Genome De Novo Sequencing and Resequencing

基因组测序工作始于20世纪70年代。1990年启动的人类基因组计划标志着基因组测序的革命性发展，在人类基因组计划开展过程中开发出的一系列关键技术，如物理图谱的构建、序列拼接、海量序列数据存储与分析等，为其他生物基因组测序计划的顺利完成提供了重要的支撑。至今，已经有较完整的全基因组序列数据的物种包括超过39种类病毒、2115种病毒、58种古细菌、1269种细菌，69种真菌、29种原生生物、10种植物和78种动物（图6-1）。随着第二代测序技术的迅猛发展，生物科学界也开始越来越多地应用第二代测序技术来解决生物学问题。比如在基因组水平上对还没有参考序列的物种进行从头测序（de novo sequencing），获得该物种的参考序列，为后续研究和分子育种奠定基础；对有参考序列的物种，进行全基因组重测序（resequencing），在全基因组水平上扫描并检测突变位点，是发现个体差异的分子基础。在转录组水平上进行全转录组测序（whole transcriptome resequencing），从而开展可变剪接、编码序列单核苷酸多态性（cSNP）、等位特异表达等研究；或者进行小分子RNA测序（small RNA sequencing），通过分离特定大小的RNA分子进行测序，从而发现新的microRNA分子。在转录组水平上，与染色质免疫共沉淀（ChIP）和甲基化DNA免疫共沉淀（MeDIP）技术相结合，从而检测出可能与特定转录因子结合的DNA区域和基因组上的甲基化位点。

1977年化学家sanger发明了双脱氧链终止DNA测序技术，并因此获得1980年的诺贝尔化学奖。这项技术一直沿用至今，被应用于基因研究的各个领域。为人类基因组计划（HGP）的完成立下了汗马功劳。

测序是根据核苷酸在某一固定的点开始，随机在某一个特定的碱基处终止，并且在每个碱基后面进行荧光标记，产生以A、T、C、G结束的四组不同长度的一系列核苷酸，然后在尿素变性的PAGE胶上电泳进行检测，从而获得可见的DNA碱基序列。Sanger法测序的原理就是，每个反应含有所有四种脱氧核苷酸三磷酸（dNTP）使之扩增，并混入限量的一种不同的双脱氧核苷三磷酸（ddNTP）使之终止。由于ddNTP缺乏延伸所需要的3’-OH基团，使延长的寡聚核苷酸选择性地在G、A、T或C处终止，终止点由反应中相应的双脱氧而定。每一种dNTPs和ddNTPs的相对浓度可以调整，使反应得到一组长几个至千以上个，相差一个碱基一系列片段。它们具有共同的起始点，但终止在不同的核苷酸上，可通过高分辨率变性凝胶电泳分离大小不同的片段，凝胶处理后可用X-光胶片放射自显影或非放射性核素标记进行检测。


2008 Genome Atlas

物种	测序中	草图	完成图	草图+完成图
类病毒			39	39
病毒			2, 115	2, 115
古细菌	32	6	52	58
细菌	638	563	706	1, 269
蓝藻	16	18	33	51
真菌	35	59	10	69
原生生物	24	23	6	29
植物	33	8	2	10
绿藻	6	2	0	2
动物	77	74	4	78
总数	861	753	2, 967	3, 720

图6-1　基因组测序发展概况

20世纪末，测序速度与质量得到了进一步的提高。第一，平板电泳分离技术被毛细管电泳所取代；第二，通过更高程度的并行化使得同时进行测序的样本数量增加。使用毛细管替代平板凝胶取消了手工上样，降低了试剂的消耗，提升了分析的速度。另外，紧凑的毛细管电泳设备的形式更易于实现并行化，可以获得更高的通量。ABI3730测序仪和Amersham Mega-BACE分别可以在一次运行中分析96个或384个样本。这一代测序仪在人类基因组计划DNA测序的后期阶段起到了关键的作用，而且由于其在原始数据质量以及序列读长方面具有优势。加速了人类基因组计划的完成。DNA测序技术经过30多年的发展，目前已经到了第三代，三代测序技术有各自的优势。通过几十年的逐步改善，第一代测序仪的读长可以超过1000bp，原始数据的准确率可以高达99.999%，每天的数据通量可以达到600 000碱基。因此Sanger法第一代测序技术仍在广泛使用，并且对于少量的序列来说，仍是最好的选择。

第二节

新一代测序技术和工作流程

Section 2 Work Flow of Next-Generation Sequencing

高通量测序技术是对传统测序一次革命性的改变，一次对几十万到几百万条DNA分子进行序列测定，因此在有些文献中称其为下一代测序技术（next generation sequencing）足见其划时代的改变，同时高通量测序使得对一个物种的转录组和基因组进行细致全貌的分析成为可能，所以又被称为深度测序（deep sequencing）。

高通量测序可以帮助研究者跨过文库构建这一实验步骤，避免了亚克隆过程中引入的偏差。依靠后期强大的生物信息学分析能力，对照一个参比基因组（reference genome）高通量测序技术可以非常轻松完成基因组重测序（resequencing），2007年Van Orsouw等人结合改进的AFLP技术和454测序技术对玉米基因组进行了重测序，该重测序实验发现的超过75%的SNP位点能够用SNPWave技术验证，提供了一条对复杂基因组特别是含有高度重复序列的植物基因组进行多态性分析的技术路线。2008年Hillier对线虫CB4858品系进行Solexa重测序，寻找线虫基因组中的SNP位点和单位点的缺失或扩增。但是也应该看到，由于高通量测序读取长度的限制，使其在对未知基因组进行从头测序（de novo sequencing）的应用受到限制，这部分工作仍然需要传统测序（读取长度达到850碱基）的协助。但是这并不影响高通量测序技术在全基因组mRNA表达谱，microRNA表达谱，ChIP-chip以及DNA甲基化等方面的应用。

2008年Mortazavi等人对小鼠的大脑、肝脏和骨骼肌进行了RNA深度测序，这项工作展示了深度测序在转录组研究上的两大进展，表达计数和序列分析。对测得的每条序列进行计数获得每个特定转录本的表达量，是一种数码化的表达谱检测，能检测到丰度非常低的转录本。分析测得的序列，有大于90%的数据显示落在已知的外显子中，而那些在已知序列之外的信息通过数据分析展示的是从未被报道过的RNA剪切形式，3’端非翻译区，变动的启动子区域以及潜在的小RNA 前体，发现至少有3500个基因拥有不止一种剪切形式。而这些信息无论使用芯片技术还是SAGE文库测序都是无法被发现的。

高通量测序另一个被广泛应用的领域是小分子RNA或非编码RNA（ncRNA）研究。测序方法能轻易地解决芯片技术在检测小分子时遇到的技术难题（短序列，高度同源），而且小分子RNA的短序列正好配合了高通量测序的长度，使得数据“不浪费”，同时测序方法还能在实验中发现新的小分子RNA。在衣藻、斑马鱼、果蝇、线虫、人和黑猩猩中都已经成功地找到了新的小分子RNA。在线虫中获得了40万个序列，通过分析发现了18个新的小RNA分子和

一类全新的小分子RNA。

在DNA—蛋白质相互作用的研究上，染色质免疫沉淀—深度测序（ChIP-seq）实验也展示了其非常大的潜力。染色质免疫沉淀以后的DNA直接进行测序，对比ref seq可以直接获得蛋白与DNA结合的位点信息，相比ChIP-chip，ChIP-seq可以检测更小的结合区段、未知的结合位点、结合位点内的突变情况和蛋白亲合力较低的区段。

一、新一代测序法和常见的测序仪

最近市面上出现了很多新一代测序仪产品，例如454基因组测序仪、Illumina测序仪、SOLiD测序仪、Polonator测序仪以及HeliScope单分子测序仪。所有这些新型测序仪都使用了一种新的测序策略——循环芯片测序法（cyclic-array sequencing），也可将其称为"新一代测序技术或者第二代测序技术"。

所谓循环芯片测序法（图6-2），简言之就是对布满DNA样品的芯片重复进行基于DNA的聚合酶反应（模板变性、引物退火杂交及延伸）以及荧光序列读取反应。2005年，有两篇论文曾对这种方法做出过详细介绍。与传统测序法相比，循环芯片测序法具有操作更简易、费用更低廉的优势，于是很快就获得了广泛的应用。

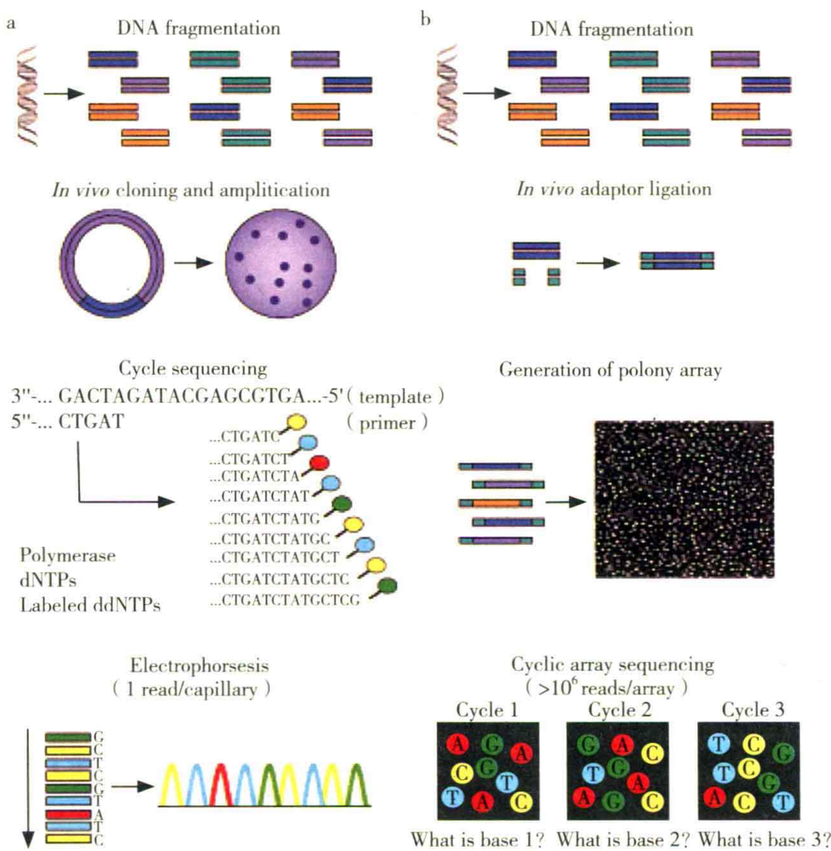


图6-2　Sanger测序法和新一代测序技术工作流程图

在开发新型高通量、高并行运行方法时碰到的一个关键问题是，如何将反应试剂同时加入数量如此之多的各个反应体系中？在焦磷酸测序的过程当中需要反复加入不同的碱基以供测序反应使用，而当时的自动化加样设备无法有效地做到对这么多的反应体系同时循环加样。于是，开发一种全新的高密度并行处理方法这一重要课题又再一次摆在了科研人员的面前。这一次，我们找到了一个非常简单但是又很巧妙地方法。在高密度的反应芯片表面使用层流（laminar flow）加样方式，反应试剂会通过扩散作用很好地进入每一个反应体系，而且也可以用层流的方式洗去多余的反应试剂。现在，所有的新一代测序仪都采用了这种层流加样方法。

为了将每个单独的测序反应都分隔开来，一开始使用平板（芯片），不过在平板上平均每一平方厘米的面积上最多只能同时进行数百至数千个反应。但我们希望达到的是在每平方厘米的面积上同时进行100万个测序反应，这样才能令测序仪小型化，同时节省试剂并进行快速成像和测序。为了实现更高密度的测序反应，我们在平板上制作了很多小孔，将每个反应体系都安置在这些小孔中，这些小孔都足够深，足以分隔每个反应体系。虽然这种方法极大提高了测序反应的密度，缩小了平板的面积，但是要达到高通量的要求还是需要60mm × 60mm大小的芯片才行。

针对图像采集问题使用了商业化的天文学照相（astrological grade camera）器材，在电荷耦合装置（CCD）的表面连接上光纤束（fiber-optic bundle）。这些光纤是锥形排列的，这样可以将大范围的光信号都传输到CCD表面上很小的一个范围。采取下面两个步骤，我们就可以制成含有高密度小孔的芯片：先将光纤束连接到类似于载玻片一样的一次性芯片上，然后用酸蚀刻（acid etching procedure）技术在玻片的另一面打上小孔。这种酸蚀刻技术是根据制作生物传感器的技术改进而来的。

二、样品准备

要想实现高通量基因组测序，只对测序步骤进行优化还是远远不够的。人类基因组计划花费经费中有很大一部分都用在了测序样品制备阶段。当时即使是采用最简单的制备样品方法也需要将目标片段克隆到细菌中，挑克隆，再转到96孔板，然后进行克隆扩增，提取质粒，制备测序模板。这种工作流程既耗时又耗钱。如果采用新型的文库制备方法就可以极大地节省这部分开支，这种新型的方法是先分离基因组DNA，随机切割成小片段分子，然后通过有限稀释（limiting dilution）和聚合酶扩增反应，即体外克隆方式（clones without bacterial）制备模板片段。这样，从模板制备到最后的测序反应整个过程都能够在体外完成。

文库制备包括以下几个步骤，首先随机切割样品基因组，获得大量DNA片段，然后接上接头进行扩增反应。新一代测序技术的样品制备程序和Craig Venter等人的鸟枪法样品制备程序有着本质的差别。通过乳糜PCR（emulsion PCR）或桥式PCR（bridge PCR）等方法对文库进行扩增，获得测序模板，而没有鸟枪法中的细菌克隆繁殖步骤。去掉了细菌繁殖步骤极大地提高了整个测序工作的速度和效率，同时避免了由于细菌繁殖导致的序列丢失的可能性。末端配对文库制备方法的建立同样对复杂基因组从头测序、对重复片段测序以及对基因组结构（复制、重排）展开系统研究三种能力。这种末端配对文库的制备方法是受到了Bender科研小组对果蝇（Drosophila）制备跨步文库方法的启发而发展得来的。

emPCR被454测序仪和SOLiD测序仪等采用（图6–3）。这种方法是将制备的DNA文库与水油包被的直径大约28 μm的磁珠在一起孵育、退火，由于磁珠表面含有与接头互补的寡聚核苷酸序列，因此ssDNA会特异地连接到磁珠上。同时孵育体系中含有PCR反应试剂，因此可以保证每一个与磁珠结合的小片段都会在各自的孵育体系内独立扩增，扩增产物仍可以结合到磁珠上。反应完成后，破坏孵育体系并富集带有DNA的磁珠。经过扩增反应，每一个小片段都将被扩增大约100万倍，从而达到下一步测序反应所需的模板量。

在桥式PCR反应中（图6–3），正向引物和反向引物都被通过一个柔性接头（flexible linker）固定在固相载体（solid substrate）上。经过PCR反应，所有的模板扩增产物就都被固定到了芯片上固定的位置。值得注意的是，Illumina测序仪使用的桥式PCR与传统的桥式PCR有所不同，它会交替使用Bst聚合酶进行延伸反应以及使用甲酰胺（formamide）进行变性反应。这样，经过桥式PCR扩增之后，也会在固相载体上形成一个个的模板"克隆"。一块芯片的8条独立"泳道"上每一条泳道都可以容纳数百万的模板"克隆"，这样一次就可以同时对8个不同的文库进行测序。

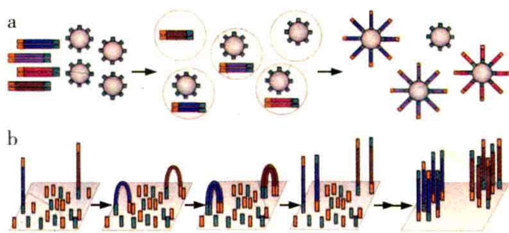


图6–3　emPCR和bridgePCR示意图

三、合成测序法

摩尔定律不仅为计算机CPU的迅猛发展提供了源动力，也给测序平台提高通量和小型化带来了希望。很明显，常规的人类基因测序项目会对我们处理测序技术的能力提出更高要求，这与我们对计算机处理能力的要求是一样的。不过，只有将计算机的电子管换成晶体管，才为后来集成电路技术的发展提供了可能，这正是计算机产业发展的关键所在。而希望对传统的毛细管电泳技术进行改良，提高它的速度和处理规模，正如只用电子管直接制作集成电路一样不可能。因此，如果将各种测序技术比作一个个晶体管，将一系列测序步骤整合起来比作集成电路，那么也就可以用摩尔定律来预测DNA测序技术的发展速度了。

合成测序法概念虽然在提出的时候还不算成功，但它的出现为测序仪小型化奠定了基础。基于合成测序法出现了两种策略：一种是循环可切除终止测序法（cyclic reversible termination technology），即依次逐个添加荧光标记的碱基，继而检测荧光信号，切除荧光基团，如此往复；另一种策略是焦磷酸测序法（sequenced by detecting pyrophosphate release）。454测序仪采用的是小型化焦磷酸测序反应，测序模板准备和焦磷酸测序反应步骤都是在固态芯片上完成的。

实际上，早在20世纪90年代中期，焦磷酸测序技术就已经被科研界用来进行基因分型工作了，但那时的焦磷酸测序技术还不能够满足标准的测序实验要求，因为它的测序长度太

短，因此只能用于旨在发现SNP的基因分型研究当中。那时焦磷酸测序还不能用于从头测序工作，因为从头测序需要对每一个尤其是第一个碱基都能准确地区分清楚，而焦磷酸测序只能简单地对已知位点的碱基进行检测，而且从头测序要求的测序长度也是焦磷酸测序法无法达到的。不过，由于焦磷酸测序的原理是通过检测碱基掺入时发出的光来进行测序的，所以它并不需要类似于电泳之类的物理分离过程来对碱基进行区分。这也就是说焦磷酸测序仪可以"缩小(减)"到只需要检测光线就够了，而不需要像传统的测序仪还需要电泳设备，而这正是限制传统电泳仪小型化的关键所在。发光检测方法还能够进行多路平行操作，但是直到454测序仪出现之前，还没有人这样做过，以前都是依次进行检测的。和晶体管早期的遭遇一样(当时人们也怀疑晶体管替代不了电子管)，人们同时对高密度的、用于并行焦磷酸测序的反应也充满了疑问。不过，当我们不在溶液中进行测序反应，而是将测序模板、所有的试剂(酶)都固定在平板上制成芯片之后，就获得了小型化的，能进行多路并行处理的测序仪，这就与晶体管被小型化并整合成集成电路的过程一样。此外，借助微量滴定板上一个个的小孔所达到的将不同测序反应进行分隔这一目的，也能通过在单个固相支持物上进行严密包裹(隔离)的反应来实现。在这些各自隔绝的反应体系中，链聚合反应速度和发光速度都能通过对反应试剂和产物弥散状况进行严密的控制来进行精密的调整(图6-4)。

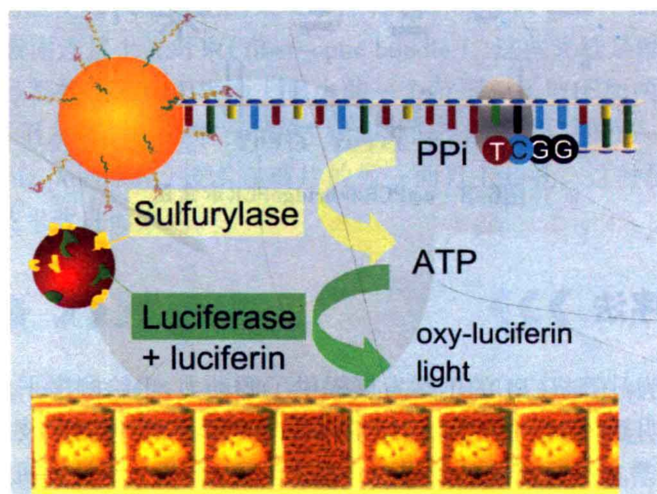


图6-4 焦磷酸测序法原理

四、第三代测序技术

近期出现的Heliscope单分子测序仪、SMRT技术和Oxford Nanopore Technologies公司正在研究的纳米孔单分子技术，被认为是第三代测序技术。与前两代技术相比，他们最大的特点是单分子测序。其中，Heliscope技术和SMRT技术利用荧光信号进行测序，而纳米孔单分子测序技术利用不同碱基产生的电信号进行测序。

Helicos公司的Heliscope单分子测序仪基于边合成边测序的思想，将待测序列随机打断成小片段并在3'末端加上Poly(A)，用末端转移酶在接头末端加上Cy3荧光标记。用小片段与表面带有寡聚Poly(T)的平板杂交。然后，加入DNA聚合酶和Cy5荧光标记的dNTP进行

DNA合成反应，每一轮反应加一种dNTP。将未参与合成的dNTP和DNA聚合酶洗脱，检测上一步记录的杂交位置上是否有荧光信号，如果有则说明该位置上结合了所加入的这种dNTP。用化学试剂去掉荧光标记，以便进行下一轮反应。经过不断地重复合成、洗脱、成像、淬灭过程完成测序。Heliscope的读取长度约为30~35 bp，每个循环的数据产出量为21~28 Gb。值得注意的是，在测序完成前，各小片段的测序进度不同。此外，可以通过二次测序来提高Heliscope的准确度，即在第一次测序完成后，通过变性和洗脱移除3'末端带有Poly（A）的模板链，而第一次合成的链由于5'末端上有固定在平板上的寡聚Poly（T），因而不会被洗脱掉。第二次测序以第一次合成的链为模板，对其反义链进行测序。

Pacific Biosciences公司的SMRT技术基于边合成边测序的思想（图6-5），以SMRT芯片为测序载体进行测序反应。SMRT芯片是一种带有很多ZMW（zero-mode waveguides）孔的厚度为100 nm的金属片。将DNA聚合酶、待测序列和不同荧光标记的dNTP放入ZMW孔的底部，进行合成反应。与其他技术不同的是，荧光标记的位置是磷酸基团而不是碱基。当一个dNTP被添加到合成链上的同时，它会进入ZMW孔的荧光信号检测区并在激光束的激发下发出荧光，根据荧光的种类就可以判定dNTP的种类。此外由于dNTP在荧光信号检测区停留的时间（毫秒级）与它进入和离开的时间（微秒级）相比会很长，所以信号强度会很大。其他未参与合成的dNTP由于没进入荧光型号检测区而不会发出荧光。在下一个dNTP被添加到合成链之前，这个dNTP的磷酸基团会被氟聚合物（fluoropolymer）切割并释放，荧光分子离开荧光信号检测区。SMRT技术的测序速度很快，利用这种技术测序速度可以达到每秒10个dNTP。

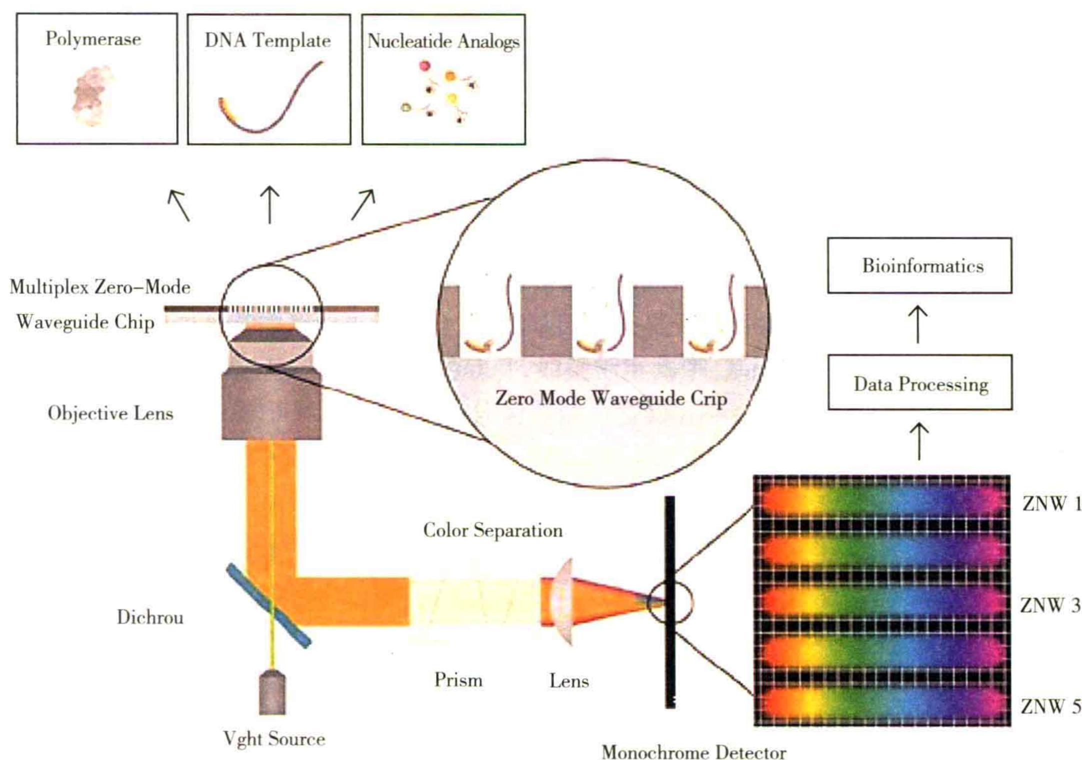


图6-5　SMRT测序技术流程

Oxford Nanopore Technologies公司正在研究的纳米孔单分子技术是一种基于电信号测序的技术。他们设计了一种以α-溶血素为材料制作的纳米孔，在孔内共价结合有分子接头环糊精。用核酸外切酶切割ssDNA时，被切下来的单个碱基会落入纳米孔，并和纳米孔内的环糊精相互作用，短暂地影响流过纳米孔的电流强度，这种电流强度的变化幅度就成为每种碱基的特征。碱基在纳米孔内的平均停留时间是毫秒级的，它的解离速率常数与电压有关，180 mV的电压就能够保证在电信号记录后将碱基从纳米孔中清除。纳米孔单分子技术的另一大特点是能够直接读取甲基化的胞嘧啶，而不像传统方法那样必须要用重亚硫酸盐（bisulfite）处理，这对于在基因组水平研究表观遗传相关现象提供了巨大的帮助。纳米孔单分子技术的准确率能达到99.8%，而且一旦发现替换错误也能较容易地更改，因为4种碱基中的2种与另外2种的电信号差异很明显，因此只需在与检测到的信号相符的2种碱基中做出判断，就可修正错误。另外由于每次只测定一个核苷酸，因此该方法可以很容易地解决同聚物长度的测量问题。该技术尚处于研发阶段，目前面临的两大问题是寻找合适的外切酶载体以及承载纳米孔平台的材料。

第三节

新一代测序数据存储、处理与分析

Section 3 Storage, Processing and Analysis of NGS Data

过去，研究人员使用ABI公司的3730XL毛细管电泳测序仪进行基因分析，每年至多能完成六千万碱基的测序量。随着测序技术日新月异的发展，这种情况已经成为历史。在2005年开始进行新一代测序技术开发时，Roche公司和454公司联合开发的焦磷酸测序仪的分析速度就已经达到了上述提及的ABI仪器速度的50倍之上。也就是从那时起，因基因数据过多而产生的问题凸显了出来，而且这个问题随着其他制造商开发出更多更快的测序仪而愈加严重。举个例子，ABI的新一代测序平台SOLiD单次运行，便可以分析6Gb的碱基序列；而Roche/454测序仪单次运行可以将上述结果转换成12~15个千兆字节（gigabytes）的数据信息；Illumina Genome Analyzer（GA Ⅱ）测序系统仅在两小时运行时间里，就得到10兆字节（terabytes）的信息。尽管可以为用户提供高达11.25TB的存储量，但对于多数实验室所具有的信息管理系统来说，规模如此庞大的数据信息，就好像是迎面而来的洪水，让人感到难以控制。

海量信息所带来的一个问题是，用户无法将初始图像数据进行分类存档，而必须利用软件对数据进行读取，然后才能对数据进行保存。对于大多数研究人员来说，像这样在每次实验后对原始数据进行处理的方式既繁琐又不经济。

除数据处理问题之外，研究人员还需要拥有一个足够强大的计算机平台，以便将来自多个测序技术的短小基因片段进行组合，形成基因组外显子。目前问题在于，测序仪生产商仅仅提供用于某些特定基因信息分析的软件，如靶标重测序、基因表达分析、染色质免疫沉淀反应或基因组从头测序等，而并未提供任何其他类型的下游生物学信息分析软件，这就给生物信息学提出了新的问题。

一、新一代测序数据格式与质量编码

目前，序列质量评分问题是受到广泛关注的一个问题。造成这种现象的原因主要是因为所有新一代测序仪的测序质量都不高，而且不同的序列情况都有各自的误差率。随着新一代测序仪产品的不断成熟，在临床及科研工作中的应用范围越来越广，它们的测序质量也就变得重要起来，而且我们也需要对各个测序仪的测序质量有一个清晰的、可靠的评价标准。对于测序仪的应用范围进行标准化的质量评价也是有好处的。比如评价从头测序的质

量、评价测序结果与参考序列的相似度、评价测序仪发现突变以及多态性的能力以及对测序仪在进行大规模测序项目研究时质量的可靠性进行评价等。

二、新一代测序数据库与数据格式转化

目前对于如何组织、存档以及发布这些新一代测序仪产生的短片段序列结果正处于热烈的讨论之中，人们希望制定一个类似芯片实验时制定的MIAME（minimum information about a microarray experiment）规则。这些早期的工作经验在如何处理包括生物学注释信息、临床原始数据、关键实验细节（比如样品特征、样品处理方法）在内的元数据，以及如何处理、出版发行这些数据等方面给了我们良好的建议。如何对这些新一代测序仪的测序结果数据进行公共管理也是一个需要探讨的问题。NCBI最近专门为短片段序列建立了数据库Short Read Archive（SRA），并同步制定数据提交格式。SRA数据库不仅会收集包括实验注释信息、实验参数等信息的数据，而且还会被整合到Entrez查询系统当中。目前的工作主要包括开发线上搜索工具、数据图形化工具。

三、测序短片段在参考基因组中的定位

新一代测序仪可以用极快的速度和极其低廉的价格获得大量的序列，这已经改变了基因组学的面貌。它们可用于对RNA进行测序，即先通过反转录将其变成cDNA，然后再对cDNA进行测序，这样就能发现一些未知的基因，并据此发现新的RNA剪切方式。也可以将测序技术应用于ChIP，弄清楚与蛋白质共沉淀的DNA片段的序列。这种方法能用于研究转录因子与DNA调控元件之间的相互作用。此外，对肿瘤细胞全基因组测序也能发现一些新的致癌突变。随着新一代测序的完成，人们获得了大量的短片段序列，如何对这些短片段作图就成了一个大问题，即被称为“阅读片段作图（‘read mapping’）”的问题。

```
@IL26_1184:6:1:881:704/1
TTTATTTTGATGCACGCACGAGACGGTATCTAGACT
+
>>>>>>>>>>><>>>>>>>>>>>>>>>>>>><<><>
@IL26_1184:6:1:883:595/1
TGGTGATTAGTCAAAGAGACCAAATCCCATATCCTC
+
>>>>>>>>>>>>>><>>>>>>>>>>>>>>><>>>><
```

图6-6 FASTQ格式示例

为了便于测序数据的发布和共享，高通量测序数据以 FASTQ 格式来记录所测的碱基读段和质量分数，如图6-6所示，FASTQ格式以测序读段为单位存储，每条读段占 4 行，第一行是这个read的编号，第二行是read的具体编码，第三行是一个标识符，第四行是read中每一个碱基对应的质量读数（这是由于每个碱基都是由测序仪测出这些质量度数是由ASCII编码）。

很多公司开发的测序仪在测序时产生的都是长约25~100bp左右的小片段序列，即“read”。这些小片段都是待测样品大片段的某一部分。与对未知的全基因组进行测序，即与

将所有小片段组装成一个完整基因组的工作相比，人们现在大部分的工作实际都可以参照"参考基因组"进行。因此，要了解小片段"read"的作用，首先要知道它们在参考基因组中的确切位置，而对这些小片段进行定位的过程就称作"作图"（mapping），或"定位"（aligning）到参考基因组中。在作图中，有一个问题需要注意，那就是进行定位时不能出现大的"间隙"。而在对RNA进行测序时，因为存在内含子的缘故，这一点就显得尤为突出。因此，对RNA进行测序时就允许有较大的间隙出现。此外，如果某个短小片段属于参考基因组里的一个重复元件，那么就应该弄清楚它来自重复元件中的哪一个拷贝。但这是不太可能实现的，所以分析程序一般都只能给出该短片段可能属于参考基因组中哪几个位点。同时，由于测序错误或者检测样品间以及检测样品和参考基因组间出现变异等情况，使上述问题变得更加严重。同样，在RNA剪切体作图中也存在上述问题，而且由于内含子的问题使得情况更为复杂。

当然，使用传统的BLAST或BLAT软件分析ChIP-seq或RNA-seq测序结果，可能会花上几百甚至几千个小时，现在有了新的分析软件。

众多测序仪每一轮测序都能获得百万计的短片段序列，不过要对一个基因组进行完全测序则需要进行好几轮检测，这也就意味着要想获得一份完整的全基因组图谱必须对数百万甚至是数十亿的短小片段进行作图、定位和拼接。比如，最近做出的癌症基因组序列就是通过132轮测序，对80亿条短小片段进行作图后得到的结果。使用BLAST或BLAT比对法，借助大型的超级计算机需要几天就能获得这个癌症的基因组序列结果，但这并非人人都能享有。为了能让更多的人用更廉价的计算机也能进行类似的作图分析，人们开发了一套新的比对定位程序，使用这种新程序即使在普通的台式机上也能对数亿计的短小片段进行作图分析。测序仪器生产厂商也会提供一些专门的作图软件，例如，Illumina公司开发的ELAND程序等。研究人员也开发了一些有针对性的第三方软件，这些软件中很大一部分都是开放源代码的免费程序。这些软件主要都是建立在这样一种算法之上，即充分利用短小DNA序列的特点来作图，而不需要依靠计算机强大的处理能力、内存容量等条件。

四、短片段作图软件

Maq和Bowtie都属于短片段作图程序（图6-7）。它们使用的是一种称作"建立索引（indexing）"的策略。同时，人们也对大量的DNA序列建立了一份索引，借助这份索引就能快速地找到其中的短DNA片段了。Maq软件是基于一种直接的但是很有效的策略——空位种子片段索引法（spaced seed indexing）。它将一个短片段（read）分成了4条长度相等的更短的片段——种子片段（seed）。如果整段短小片段（read）可以与参考基因组序列完全配对，那么很显然所有的种子片段（seed）也理所应当地应该与参考基因组序列完全配对。但如果其中有一处错配，例如SNP，那么肯定有一条种子片段无法与参考基因组序列完全匹配。以此类推，如果出现了两处错配就会导致一条或两条种子片段无法与参考基因组序列完全匹配。因此，对所有种子片段两两组合后的片段（共有6种组合方式）进行比对，就有可能找出该短小片段在基因组中最有可能的位点。Maq软件采用的这种"空位种子片段索引法"（spaced seed indexing）作图时的效率非常高。

Bowtie软件采用的则是另一种完全不同的策略，该策略借鉴了Burrows- Wheeler转换（Burrows-Wheeler transform）这种数据压缩算法技术，将完整的人类基因组序列索引压缩

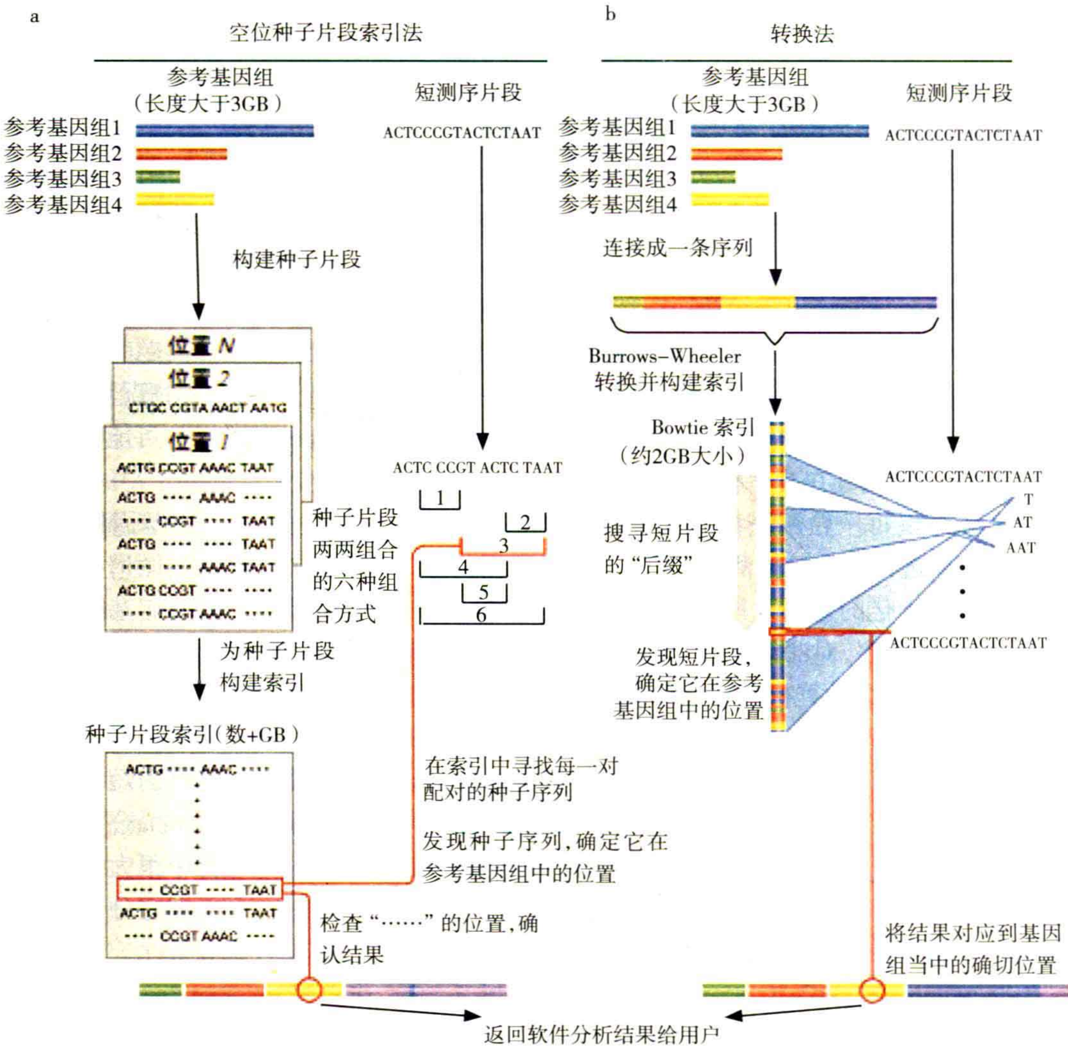


图6-7 两种短片段定位方法

到不到2GB大小（这是当前主流台式机甚至是笔记本电脑都能达到的水平），而空位种子片段索引法至少需要50GB。Bowtie每次都只把一段短片段序列中的一个碱基与经Burrows-Wheeler转换压缩过的参考基因组序列进行比对。经过这种连续的比对，最终也能找出这段短片段在参考基因组中的定位。如果Bowtie软件发现短片段中的某个碱基在参考基因组中没有很好地配对，那么软件就会退回到上一个碱基重新进行比对。实际上，Burrows-Wheeler转换使得Bowtie软件通过碱基逐个比对，直至完成全长短序列比对的方法解决了短序列作图的问题。从本质上来说，Bowtie软件使用的算法要比Maq采用的复杂得多，但Bowtie软件却比Maq软件分析的速度快30倍。

Bowtie软件和Maq软件的默认模式中至多都只会允许两个错配位点，不过有时有些用户需要允许更多的错配位点存在。还有一些测序项目，例如细菌或真菌基因组测序项目等获得的片段序列与目前已经测得的类似物种全基因组序列之间存在着较大的差异。再加之随着新测序仪的不断涌现，测序结果的质量也在不断提高，但这些测序结果却极易受到各种因

素的影响，例如样品文库的准备、测序操作步骤、甚至是放置测序仪器实验室的温度等。鉴于此，面对上述这些新出现的"问题"，人们也应该采取相应的措施，调整Maq软件和Bowtie软件的各种参数使之适应这些新情况。

Bowtie软件包中包括预置的大肠埃希菌基因组索引和部分大肠埃希菌短片段序列。要使用该软件分析数据只需输入命令就会生成一个表格式的报告，给出每一个匹配短序列的编号、在参考基因组中的位置以及发生错配的位点个数和具体位置。

有了序列定位的软件，接下来就可以了解这些短片段具体在参考基因组中的什么位置了，同时也可知道SNP都位于基因组中的什么地方。SAM软件包能满足这些要求。SAM软件包（http: //samtools.sourceforge.net）包括一体化的碱基调用和浏览器（base caller and viewer），它能使用Maq和Bowtie两种分析软件的结果。

五、基因表达水平估计

为了保持对不同基因和不同实验间估计的基因表达值的可比性，人们提出了 RPM 和 RPKM 的概念。RPM（reads per million reads）即每百万读段中来自于某基因的读段数，考虑了测序深度对读段计数的影响，RPKM（reads per kilo bases per million reads）是每百万读段中来自于某基因每千碱基长度的读段数，公式表示为：

$$RPKM=\frac{\text{基因区域read数}}{\text{基因长度}\times\text{测序深度}}\times 10^9 \qquad \text{（公式6-1）}$$

另外，对于采用末端配对测序法（paired-end sequencing）技术获得的数据，cufflinks软件等也采取了其他标准，如FPKM（fragments per kilobase of exon model per million mapped fragments）。

六、可变剪切作图软件包

要将RNA的反转录片段cDNA重新定位到基因组当中需要更加复杂的专业化算法。要将不同外显子经过剪切拼接之后生成的RNA短片段重新定位到基因组中和将一个外显子生成的RNA短片段重新定位到基因组中是完全不一样的。

在RNA反转录产物cDNA的定位操作中用到的诸如ERANGE（http: //woldlab.caltech.edu/rnaseq）这类软件包都会用到已知基因的外显子位置和内含子位置信息作为参考。这样，ERANGE软件包就能"横跨"多个外显子构建新的参考序列，然后再调用Maq程序或者Bowtie程序将剪切后的RNA片段定位到参考序列中了。因为这种方法不能发现新的（人们未知的）剪切模式，所以有些科研人员就使用了一种"机器学习法"（machine learning method）来预测新的剪切模式。该方法借助现有的参考序列注释信息在统计模型（statistical model）上进行过演练。与此相反，TopHat软件包（http: //tophat.cbcb.umd.edu）则不需要借助任何注释信息，它使用的是Bowtie软件来发现包含有短片段的外显子，然后再将余下的短片段定位到前面发现的各种外显子连接体当中。

Maq、Bowtie以及其他几种短片段作图软件都可以处理长度超过100bp的测序片段结果，但这只是在特定的情况下，而且只有原本就是针对长片段设计的软件，例如BLAT才能更好

地处理这类测序结果。另外,如果测序的样品物种序列和现有的参考序列差异很大,那该如何调整作图软件的参数呢？软件能够自动调整参数吗？这样做出来的图质量又如何呢？上述这些问题的解决方案都依赖于采用的检测方法和分析范围。不过,随着技术的进步,相信所有这些问题很快都会被攻克。

第四节

DNA和RNA测序

Section 4 DNA and RNA-Seq

DNA-seq在疾病中已经得到了广泛的应用，Shusuke Akamatsu等对2557个前列腺癌症样本和3003个正常样本的DNA-seq结果进行关联分析，发现11q12、10q26和 3p11.2这几个区域和前列腺癌易感有显著关联。Chizu Tanikawa等在日本人群中发现了两个和十二指肠溃疡易感显著关联的位点。Sun等人使用RNA-seq等测序方法发现在前列腺癌组织中发现可能导致癌症融合基因。新一代测序技术除了在疾病中的常规应用外，其他方面的使用前景也很好。

一、DNA重测序与个体变异发现 》》》

人类基因组上广泛存在着多种遗传变异形式与DNA多态性。单个核苷酸的变异早已被熟知，其中那些频率大于1%的被称为单核苷酸多态性（SNP）。国际人类基因组单体型图计划（international HapMap project）已经在人类群体中发现了数百万计的SNP。尽管一部分的SNP被发现与人类疾病相关，但只能解释疾病遗传因素中的一小部分，仍有较多的未知遗传因素（missing heritability）没有被揭示。2008年初启动的“千人基因组”计划由来自英国桑格研究所，美国国立人类基因组研究所，中国深圳华大基因研究院等多家机构共同完成。在这一计划中，科学家们对全球各地至少1000个（目前是2000个人左右）人类个体的基因组进行测序，寻找基因与人类疾病间的秘密关系。通过这些测序也将生成一个庞大的、公开的人类基因变异目录，有助于进行分析以及个体化医疗。千人基因组计划完成并公布了首项研究成果，包括对三个人群的179人按低覆盖率进行全基因组测序；对两个由“母亲-父亲-孩子”组成的三人组按高覆盖率进行测序；对来自七个人群的697人进行以外显子为目标的测序。这项研究找出了1000多万个大大小小的基因变种，其中约800万个都是以前所未知的。对于人群携带率在1%以上的基因变种，本次研究的覆盖率达到95%以上。这一成果在医学等领域有很高的应用价值，比如通过参照图谱，可以方便地找出致病的基因变种。另外研究人员还验证了在大型基因研究中综合使用多种基因测序手段的可行性。由于基因测序成本目前仍很高昂，如果能在“精测”一些基因序列的同时，对另一些基因序列只需“粗测”就能保证最终结果的准确性，将可以大幅降低基因测序研究的成本。Science相关文章对这一方面进行了介绍，文中提到研究人员开发出了几种分析和计算技术克服了对多拷贝基因进行研究的障碍，利用这一新方法，研究人员对1900个碱基对长的DNA片段拷贝数进行精确估

计，拷贝数的计数范围为0~48之间。

除了DNA的点突变，基因组上还可以发生涉及大片段DNA序列的变异，包括亚显微结构（sub-microscopic）的微重复（microduplication）和微缺失（microdeletion）。此类基因组片段的拷贝数变异（copy number variation，CNV）和SNP类似，除了一部分会致病以外，也可以作为一种遗传多态性存在于人类及其他物种的基因组上。有两个研究小组借助于新一代测序技术，几乎同时发现了人类基因组中CNV广泛分布，不仅作为一种遗传多态性在人类基因组中广泛分布，而且可以导致出生缺陷、对艾滋病病毒的易感性、对孤独症和精神分裂症的易感性等复杂疾病。已经报道的基因组结构变异（structural variation，SV）超过66 000个，其中主要是CNV。借助于新一代测序技术和相应的实验策略，如paired-end mapping（PEM）与基于测序深度（Read depth）检测的分析方法，对CNV进行高通量无偏差的发现和精确定位。人类基因组结构变异研究组（human genome structural variation group）和千人基因组计划已经获得了初步数据，包括1500万个SNP，100万个短的插入或缺失以及2万个CNV的位点，其中绝大部分都是新的发现。

二、细菌基因组测序与致病性位点发现

一个合作研究项目采用454测序仪对4株结核分枝杆菌基因组进行测序，这四株结核分枝杆菌分别是一株对R207910具有耐药性的结核分枝杆菌（mycobacterium tuberculosis）菌株，基因组大小约4Mb；两株对R207910具有耐药性的耻垢分枝杆菌（mycobacterium smegmatis），基因组大小约6Mb；以及一株正常的耻垢分枝杆菌，基因组大小约6Mb。他们希望能发现结核分枝杆菌对R207910产生抗药性的机制。该项研究在只有一位实验人员参与实验的情况下，包括样品制备等步骤在内所用的时间仅需要一周，而且避免了传统测序方法中细菌克隆阶段可能出现的错误，获得了高质量的测序结果，发现了导致结核分枝杆菌对R207910产生抗药性的两个点突变位点。这项研究成果让我们在最近的40年内第一次找到了特异性治疗结核病的药物。随后研究人员开展了一系列采用新一代测序仪的研究项目，对高致病性细菌空肠弯曲菌（campylobacter jejun）基因组的从头测序项目、对幽门螺杆菌（helicobacter pylori）在慢性胃炎致病过程中的进化研究项目、从南极海冰细菌（Antarctic sea ice bacterium）中新发现冰结合蛋白（ice-binding protein）并对其测序的研究项目，以及在引起肺炎、脑膜炎和泌尿道感染的细菌中发现致病因素的研究项目等。

三、宏基因组测序与感染性疾病分析

美国在2001年暴发了炭疽恐怖袭击危机之后，研究人员开始针对复杂的、未知的、未人工培养的环境微生物基因组进行测序。在一个研究项目中，有三名患者都接受了同一名澳大利亚器官捐赠者的器官，之后均因不明原因而死亡。从这三名死者身上提取了非人类DNA样品进行测序，结果获得了144 000条序列。分析后发现，这些序列分别属于一种沙粒病毒科（Arenaviridae）家族病毒的14个不同基因。随后进行的第二项研究在对健康蜂群和患病蜂群进行环境基因组学比较研究之后发现，以色列急性麻痹病毒（Israeli acute paralysis virus）是导致蜜蜂蜂群崩溃症的元凶。这些研究都突出了新一代测序仪的一个特点，即在样

品准备前不需要进行克隆或预扩增步骤，因此非常适用于对未知的未能人工培养的物种进行测序。这些特点也在其他对地下矿藏、深海、土壤和高盐等环境下进行的环境微生物构成方面的研究所证实。

四、古生物基因组和进化研究

要用传统的测序方法对尼安德特人的基因组进行测序研究非常困难，因为这些古老DNA量非常少，而且都早已裂解成了片段。一个国际性的研究团队对尼安德特人（Neandertal）的基因组序列进行了测定。他们所用的是在克罗地亚的一个洞穴中发现的来自3个尼安德特人骨头的一个药片大小的骨粉样品。他们将这些尼安德特人的基因组与来自世界不同地区的5个现代人的基因组进行了比较。结果显示，人类拥有多种独特的基因，其中包括在人类与尼安德特人从一个共同祖先分开之后少数在我们的人类种系中快速扩散的基因。研究还发现在人类中经常发生但在尼安德特人中却不发生的基因序列变异的基因组区域。他们找到了212个有这种变异的区域。在其中20个区域中，有着最强的正向选择证据的是3个基因，当它们发生突变的时候，可影响思维和认知能力的发展。这些基因被认为与Down综合征、精神分裂症和自闭症有关。该团队的带头人Pääbo说："获得第一个版本的尼安德特人的基因组测序完成了人们的一个长期以来的梦想。我们第一次能够发现将我们与其他所有生物区别开来的基因特征，其中包括那些在进化上距离我们最近的亲族。"尼安德特人第一次出现的时间大约在40万年之前，其分布遍及欧洲和西亚，并在大约3万年前灭绝。Pääbo带领的另一项研究提出了对尼安德特人基因组的选择区域（特别是那些来自已经降解的尼安德特人遗骨）进行测序的新技术。他们应用一种"目标序列捕捉"的方法来加强他们对来自西班牙的另外一个尼安德特人个体的基因组中的数个片段的蛋白编码区域的聚焦。他们发现了88个替代氨基酸，这些氨基酸在我们与尼安德特人分开之后已经成为固定的状态。尼安德特人的基因组片段长度基本上都介于40~90bp之间，而且最近开发的乳液PCR方法也能够对微量（单分子）样本进行很好的扩增。

五、外显子组测序

外显子组是指全部外显子区域的集合，该区域包含合成蛋白质所需要的重要信息，涵盖了与个体表型相关的大部分功能性变异。外显子组序列捕获及第二代测序是一种新型的基因组分析技术。与全基因组重测序相比，外显子组测序只需针对外显子区域的DNA即可，覆盖度更深、数据准确性更高，更加简便、经济、高效。可用于寻找复杂疾病如癌症、糖尿病、肥胖症的致病基因和易感基因等的研究。目前许多科学家都利用这一方法找到了致病基因，比如美国国家心肺血液研究所就从4名弗里曼谢尔登综合征患者的DNA中准确找出了致病基因变异。他们的研究表明，对于单个基因变异引起的疾病，外显子测序同样可以准确找到致病基因，与全基因组测序无异。研究人员认为，外显子测序也可用于多重基因变异引起的常见疾病，如糖尿病和癌症的研究中，来揭示该种疾病的致病基因。

来自华盛顿大学医学院的研究人员利用外显子组测序方法，找到了一种致命性眼睛癌症的关键基因，这一研究成果可能作为未来治疗这种癌症的靶标，并且用于其他具有高度转

移性癌症的治疗靶标。葡萄膜恶性黑色素瘤（maligment melanoma of uvea）是成年人中最多见的一种恶性眼内肿瘤，在国外其发病率占眼内肿瘤的首位，在国内则仅次于视网膜母细胞瘤，位列眼内肿瘤的第二位。此瘤的恶性程度高，易经血流转移，在成年人中又是比较多见，在临床工作中易与许多眼底疾病相混淆。由于这种癌症转移程度很高，因此要找到关键的基因并不容易，之前的研究发现这种癌症涉及调节蛋白降解的特别基因的缺陷，为了进一步分析葡萄膜恶性黑色素瘤，研究人员采用了外显子组测序方法，结果发现在研究人员分析的31个肿瘤样本中有26个（占84%）在一个叫做BAP1的基因中存在着失活性突变。研究结果发现，BAP1信号转导通路不但可作为葡萄膜黑色素瘤的一种治疗目标，而且它还有可能作为其他具有高度转移性的癌症的治疗目标。

六、非编码RNA测序

454测序仪具有不需要进行传统的细菌克隆步骤，而且足以覆盖只有21bp长的miRNA的测序长度等优势。其最早参与进行的miRNA研究是对拟南芥（arabidopsis thaliana）miRNA开展的研究。随后马上又参与了另一项研究项目，在这个项目中我们在小鼠体内发现了一种新型的小RNA——piRNA。这些研究项目为我们在人类、黑猩猩、斑马鱼和肿瘤细胞系中开展小RNA研究铺平了道路。454测序仪具有的这种对小RNA进行研究的能力使它在众多有关RNA的研究领域都能有所作为，例如转录体研究领域、EST研究领域研究领域和基于转录体的SNP研究领域等。

七、核糖体印记与深度测序技术

将核糖体图谱（ribosome profiling）和深度测序（deep sequencing）相结合，研究人员可以从基因组水平监测蛋白质的翻译状况。深度测序的强大功能对生物学研究的各个领域都产生了极大的影响。在诸如全基因组测序等方面，新技术的高效性和经济性使人们得以以一种以前无法想象的方式进行试验研究。而在另一些情况下，例如RNA测序时，借助深度测序可以进行更多的定量分析，获得更大的动态范围。在另一些研究中，例如最近由美国加州大学（University of California）的Jonathan Weissman小组发表的有关翻译图谱（translational profiling）的研究中报道的那样，深度测序不仅是一个有效的定量手段，同时还能提供很多有用的新信息。

使用核酸酶消化mRNA时，在翻译过程中发挥作用的核糖体结合并保护了大约30bp的mRNA片段。这些被保护的mRNA片段构建成DNA文库，再使用测序仪对文库中所有的片段进行测序，最终得到有关细胞中蛋白质翻译情况。

这种方法可以应用于很多方面。首先，它能广泛地用于蛋白质组研究当中。这种新方法用于研究酵母，因为酵母比较简单，同时也被研究得比较透彻，因此相对来说比较容易研究。但是从理论上来说，该方法是可以应用到其他任何一种物种中的。另外，将该技术与标记有抗原表位的核糖体（epitope-tagged ribosomes）结合使用，还有可能用于研究组织特异性的蛋白质翻译（tissue-specific translation）。

其次，在检测蛋白质表达情况时，使用核糖体图谱技术相比检测mRNA丰度来说更准

确。研究人员借助核糖体图谱技术为胞内数千种mRNA构建了核糖体印记密度图谱，并通过这些数据获得了蛋白质翻译表达速度方面的数据。据这些研究人员报道，使用蛋白质翻译表达速度方面的数据来判断蛋白质丰度要比用mRNA丰度来预测准确得多。实际上，如果对结合在mRNA链5’端的核糖体数目进行进一步的修正，就能更准确地预测出蛋白质的丰度。

核糖体图谱还可以用于翻译控制(translational control)分析。核糖体图谱技术具有很高的空间准确性(spatial precision)，能准确地反映出究竟是哪一个阅读框被翻译了。因此，可以使用该技术研究程序性框移(programmed frameshift)和终止密码子通读(stop-codon readthrough)等现象。

第五节

研究实例：基于新一代测序技术的癌症组学研究

Section 5 Case Studies: Canceromics Research basd on the Next Generation Sequencing Technology

一、短序列数据准备

短片段序列数据库Short Read Archive(SRA)是美国国立生物技术信息中心网站中的一个存储新一代测序数据的数据库，它提供了包括实验注释信息、实验参数等信息的测序数据，可以在该数据库中检索并下载感兴趣的数据。下面分6个步骤对下载数据进行实例操作。

步骤1：访问NCBI SRA数据库http: //www.ncbi.nlm.nih.gov/sra(图6-8)，在搜索框中输入感兴趣的查询关键词，例如“lung cancer”。

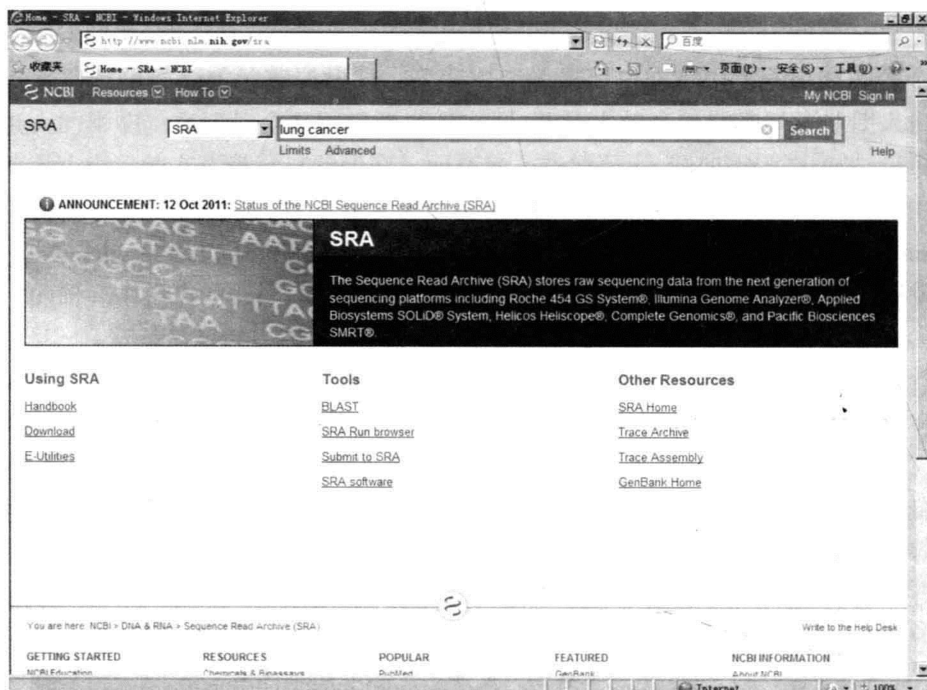


图6-8 SRA数据库网站首页

步骤2：关键词“lung cancer”得到78个查询结果（图6–9），页面右上方提供了过滤查询结果的条件，包括使用权限、数据来源和数据类型。此处按照使用权限，有两套数据需要申请才能获得，剩余的76套数据可以免费下载。

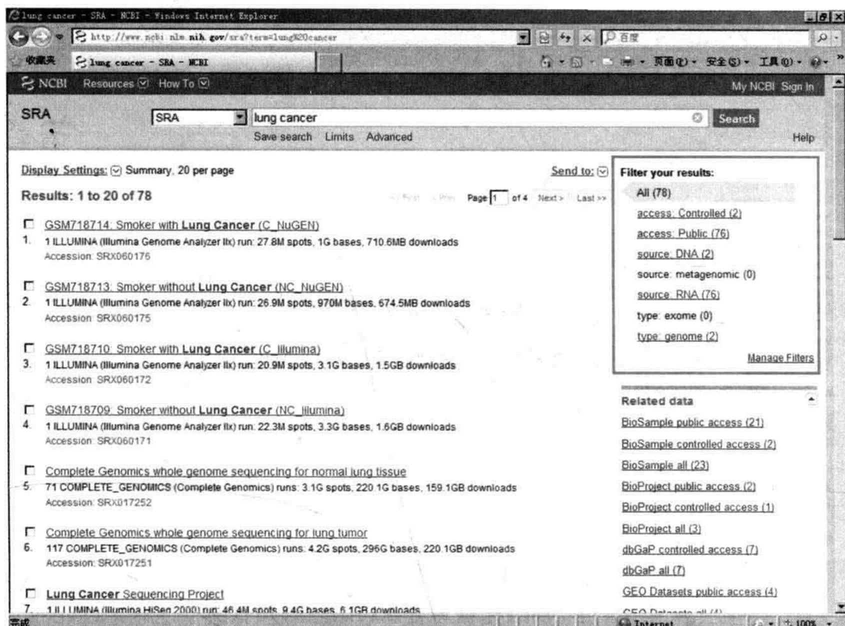


图6–9　SRA数据库查询结果

步骤3：过滤条件选择“access: Public”得到76个能够自由下载的数据列表（图6–10）。选择第一套数据GSM718714：Smoker with Lung Cancer（C_NuGEN），并查看样本信息。

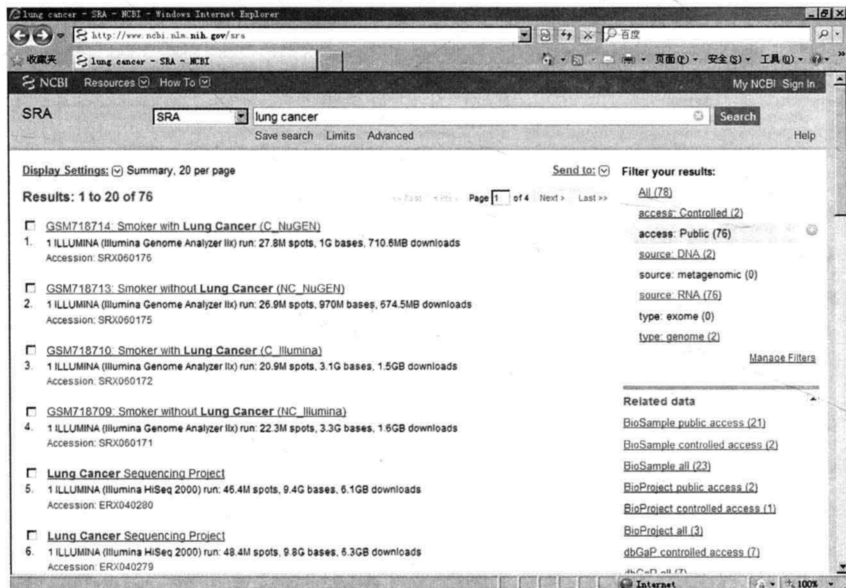


图6–10　SRA数据库查询过滤结果

步骤4：GSM718714：Smoker with Lung Cancer（C_NuGEN）的样本信息包含关于研究（Study）、样本（Sample）、实验（Library）等信息的描述（图6-11）。

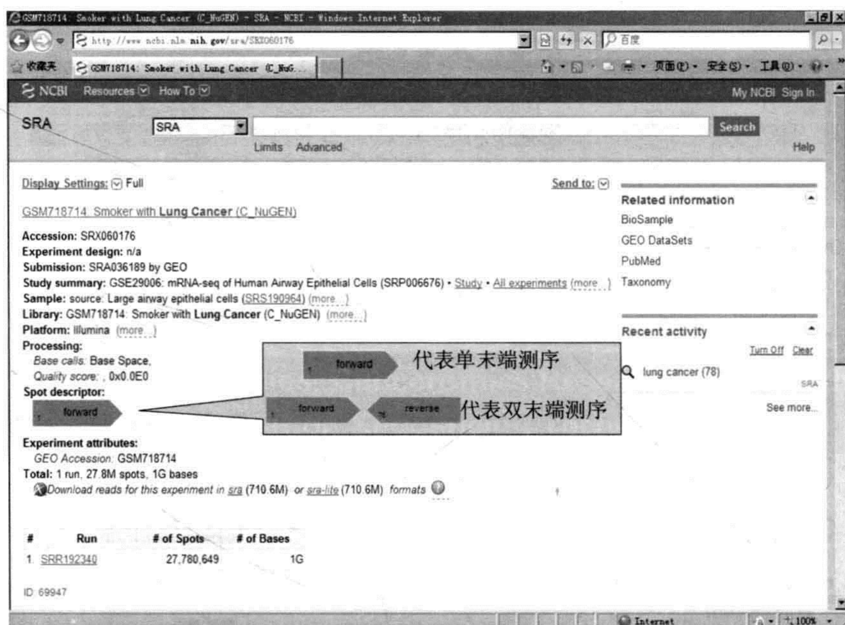


图6-11 SRA数据库样本GSM718714信息

步骤5：点击“more...”即可查看更详细的信息（图6-12）。其中Sample的详细描述包括实验细胞类型，样本性别，吸烟状态等信息；Library的详细描述包括实验平台，测序类型，测序长度等信息。

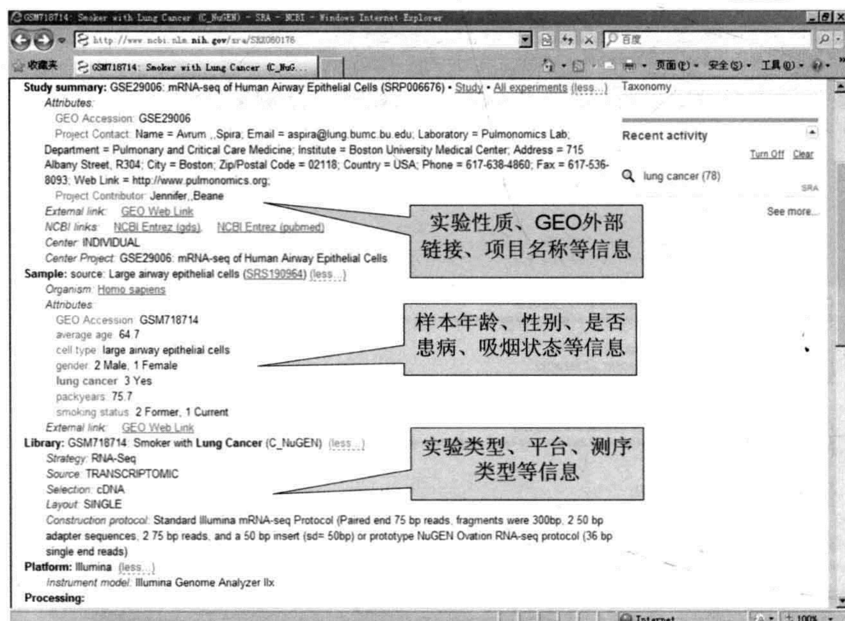


图6-12 样本GSM718714部分详细信息

步骤6：点击“Study summary”中的“Study”可以链接到样本GSM718714所在实验SRP006676的完整记录（图6–13），右上方提供该实验包含全部样本的数据下载链接，并可以选择安装Aspera plugin软件来加快下载速度，右下方提供该实验包含的单个样本的数据下载链接。

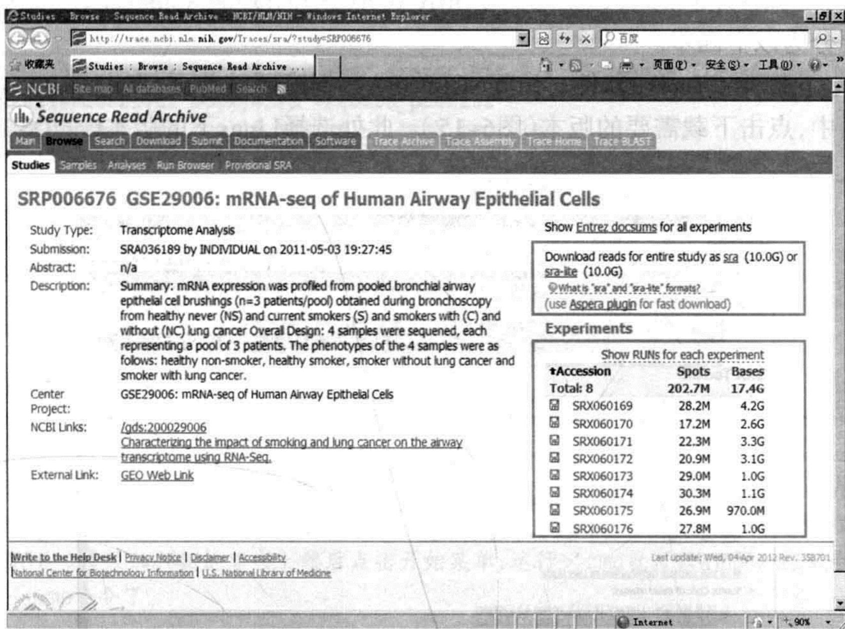


图6–13　实验SRP006676的完整记录

使用Aspera plugin下载样本单末端测序样本GSM718714的数据SRR192340，为了后面求差异表达（图6–14），另外下载两个双末端测序数据，一个是健康样本GSM718707的数据SRR192333，一个肺癌样本GSM718710的数据SRR192336。

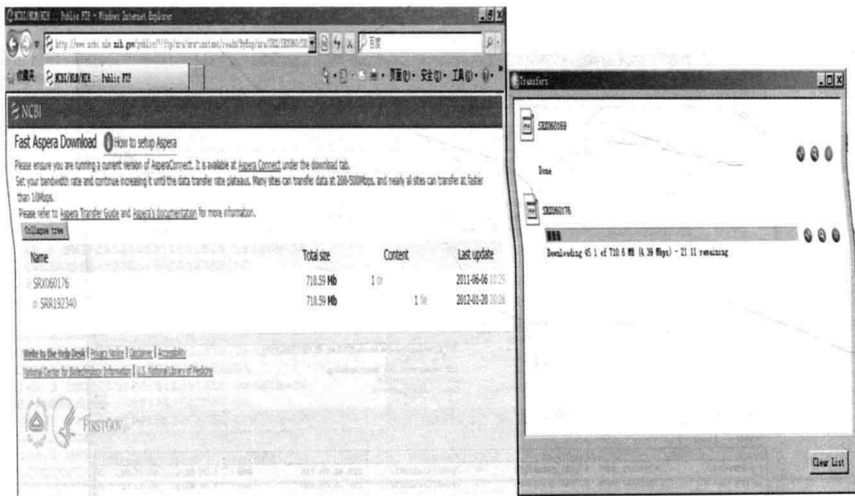


图6–14　使用Aspera plugin下载数据

二、短序列数据格式转换

下载得到的以.sra结尾的文件是一个压缩格式的文件，无法直接阅读，使用前需要通过软件将其转换为.fastq等格式，这就用到了SRA Toolkit中的fastaq-dump命令。进行短序列格式转换主要分为以下四步。

步骤1：下载SRA Toolkit软件。访问NCBI SRA数据库网站首页，点击“SRA software”，在打开的页面中，点击下载需要的版本（图6-15）。此处选择Linux下的版本CentOS Linux 64 bit architecture。

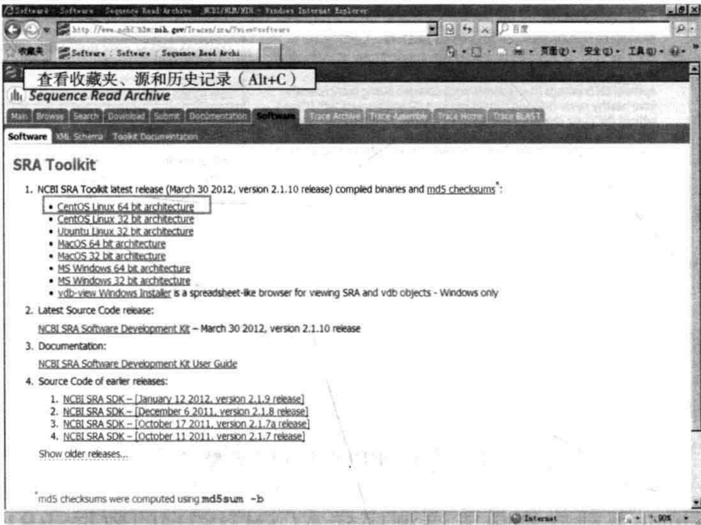


图6-15　下载SRA Toolkit

步骤2：使用Xmanager的Xftp将下载的sratoolkit.2.1.10-centos_linux64.tar.gz上传到服务器（图6-16）。

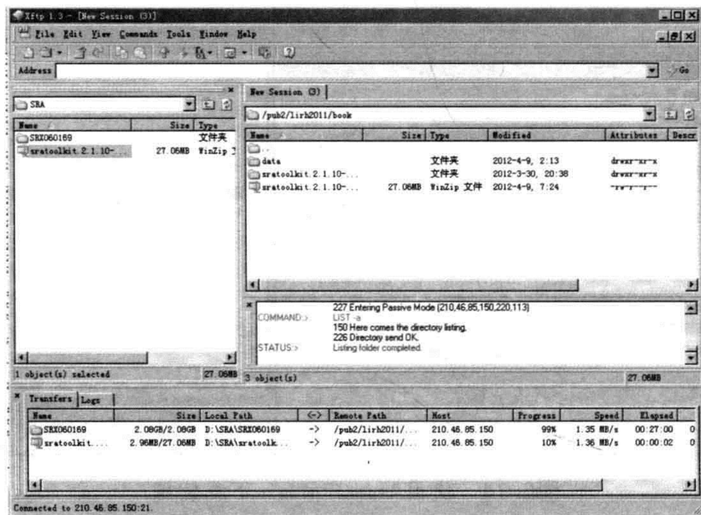


图6-16　sratoolkit压缩包上传到服务器

步骤3：将sratoolkit.2.1.10-centos_linux64.tar.gz解压缩，并将目录sratoolkit.2.1.10-centos_linux64/bin添加到linux系统的环境变量PATH中，这样在使用该程序的时候就不用输入该程序的绝对路径（在文件~/.bash_profile文件中添加export PATH= /sratoolkit.2.1.10-centos_linux64/bin: $PATH），这样用户就可以在控制台中使用sratoolkit的各项工具了，其中fastq-dump命令可以进行序列转换（图6-17）。

```
[lirh2011@webserver book]$ tar -xzf sratoolkit.2.1.10-centos_linux64.tar.gz
[lirh2011@webserver book]$ vi ~/.bash_profile
[lirh2011@webserver book]$ source ~/.bash_profile
[lirh2011@webserver book]$ fastq-dump

Usage:
  fastq-dump [options] [ -A ] <accession>
  fastq-dump [options] <path [path...]>

Use option --help for more information

fastq-dump : 2.1.12
```

-A指定输出文件名

指定要转换的.sra文件的路径

图6-17 安装SRA Toolkit

注：如果选择Windows版本，则将下载后的文件解压缩，将sratoolkit.2.1.10-win32/bin目录添加到系统的环境变量PATH中（右键桌面的我的电脑->属性->高级->环境变量->系统变量->PATH->修改加入新的路径），然后点击开始菜单，运行->cmd就可以Windows控制台中执行fastq-dump命令了。

步骤4：使用fastq-dump进行文件格式转换。对于单末端测序的数据（如SRR192340）可直接执行fastq-dump命令，输出结果为一个.fastq文件；对于双末端测序数据（如SRR192333），执行fastq-dump命令时需要添加参数--split-3来指定输出为2个文件（图6-18）。

```
[lirh2011@webserver book]$ cd /pub2/lirh2011/book/data
[lirh2011@webserver data]$ fastq-dump --split-3 -A SRR192333 /pub2/lirh2011/book/data/SRX060169/SRR192333/SRR192333.sra

Written 28216213 spots for /pub2/lirh2011/book/data/SRX060169/SRR192333/SRR192333.sra
Written 28216213 spots total
[lirh2011@webserver data]$
[lirh2011@webserver data]$ fastq-dump -A SRR192340 /pub2/lirh2011/book/data/SRX060176/SRR192340/SRR192340.sra
Written 27780649 spots for /pub2/lirh2011/book/data/SRX060176/SRR192340/SRR192340.sra
Written 27780649 spots total
[lirh2011@webserver data]$ ls
SRR192333_1.fastq  SRR192333_2.fastq  SRR192340.fastq  SRX060169  SRX060176
[lirh2011@webserver data]$
[lirh2011@webserver data]$ less SRR192340.fastq
@SRR192340.1 HWI-EAS266:3:1:0:629 length=36
CCTAGGGCCACAGACTTGCAGGTGTCTGAGCTGAGC
+SRR192340.1 HWI-EAS266:3:1:0:629 length=36
ABB?ABBBB@BAAAABA<?BBB>;@=AA?@AB@;>@
@SRR192340.2 HWI-EAS266:3:1:0:488 length=36
CCTCCGACTTTCGTTCTTGATTAATGAAAACATTCT
+SRR192340.2 HWI-EAS266:3:1:0:488 length=36
BBCCCCBBCCCBCBBACCCBCBBBBCCBBCCCCBBC
@SRR192340.3 HWI-EAS266:3:1:0:380 length=36
CCGCACCCAATTGGACCAATCTATCACCCTATAGAA
+SRR192340.3 HWI-EAS266:3:1:0:380 length=36
BBA@BBA@9>?BBB6B?><>AB@??A>@ABBB@=9A
@SRR192340.4 HWI-EAS266:3:1:0:467 length=36
AGGTGTGAGGAGTTCAGTTAAATGTTTGGGATTTTT
+SRR192340.4 HWI-EAS266:3:1:0:467 length=36
?;CACBCCCCCCCCCCCCCCCCCCCCCCCCCC>C=
```

双末端测序

单末端测序

转换后文件列表

每4行表示一条read
第1、3行是ID和长度信息
第2行是序列，第4行是质量数

图6-18 运行fastq-dump结果

三、短序列在参考基因组上定位

以下分六个步骤将转换成fastq格式的短序列文件定位到参考基因组上。

步骤1：访问bowtie软件首页http: //bowtie-bio.sourceforge.net/index.shtml（图6-19），页面右侧提供软件的使用说明、下载链接、与软件相关的其他工具的链接以及预先构建好的参考基因组。

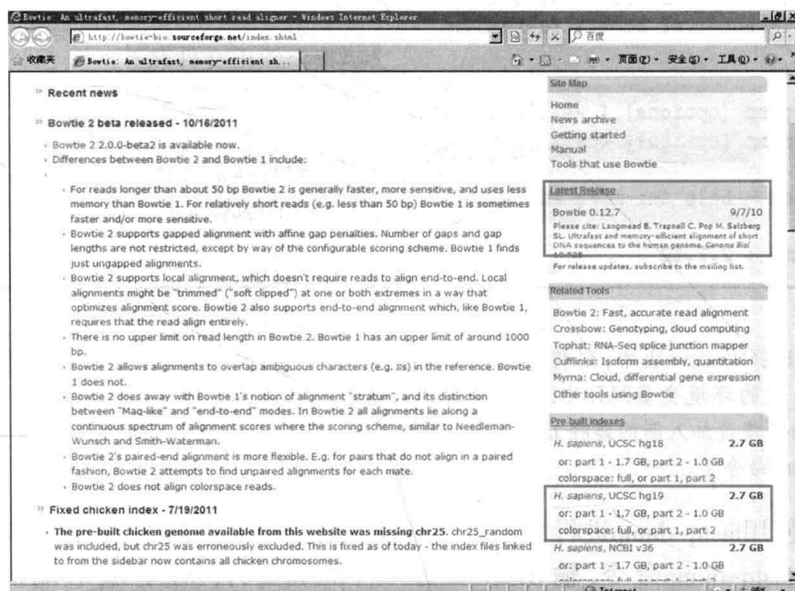


图6-19 软件bowtie首页

点击Latest Release中的Bowtie，打开下载页面（图6-20）。

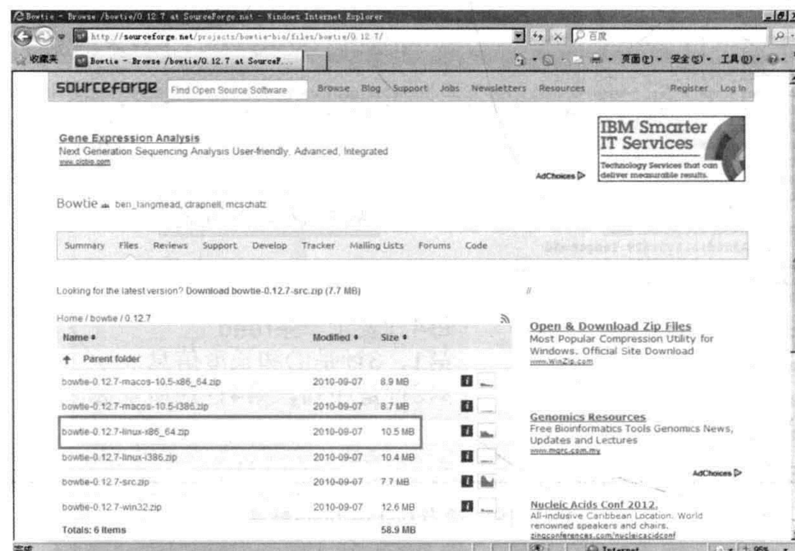


图6-20 下载bowtie的安装文件

步骤2：安装bowtie。将下载的bowtie-*.*.*-linux-x86_64.zip文件上传到服务，使用unzip命令将文件解压缩，然后将得到的目录bowtie-*.*.*添加到linux系统的环境变量PATH中（在文件~/.bash_profile文件中添加export PATH= /bowtie-*.*.*: $PATH）。

步骤3：准备参考基因组。在bowtie软件首页上点击Pre-built indexes中的H.sapiens，UCSC hg19下载人参考基因组hg19（人参考基因组hg19共2.7G，建议使用下载软件下载）（图6-21）。将下载的hg19.ebwt.zip文件上传到服务器，使用unzip将其加压缩。

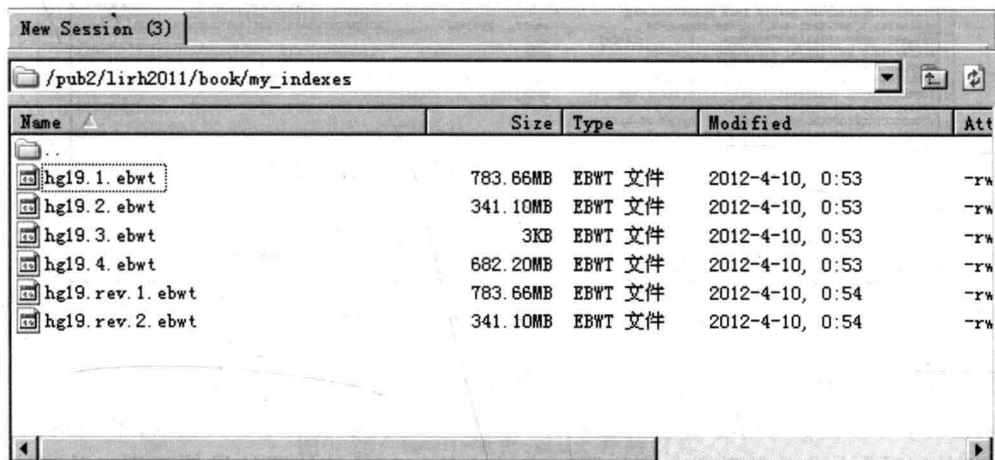


图6-21　人类hg19参考基因组

步骤4：执行bowtie命令，将短序列Read定位到参考基因组上。在控制台中输入bowtie命令，可以得到各项参数设置方式，对于单末端测序的数据SRR192340，以及双末端测序数据SRR192333执行命令及结果如下（图6-22）。

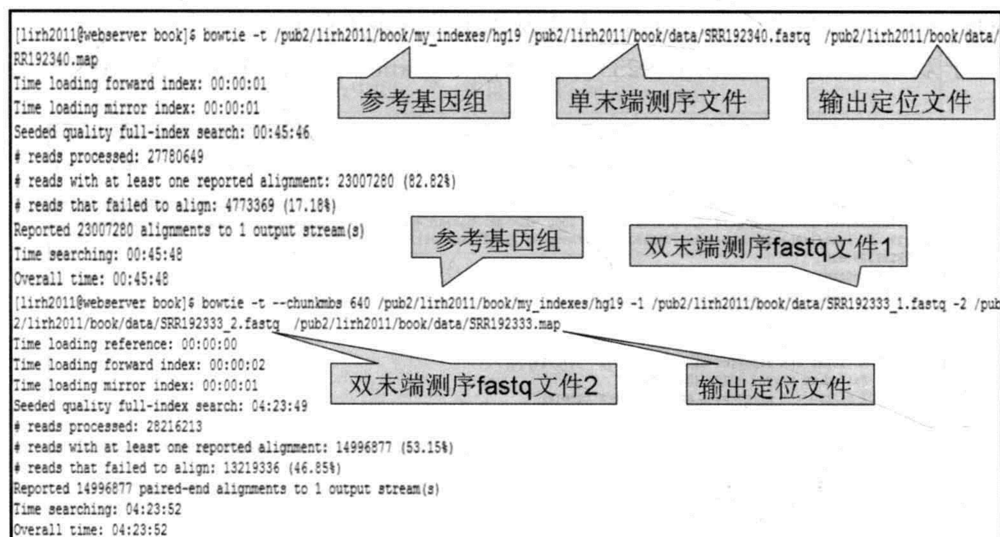


图6-22　短序列Read定位到参考基因组

其中参数-t显示运行时间，bowtie默认的最大存储空间为64MB，使用参数----chunkmbs可增大内存。

从运行结果可以看出短序列匹配到参考基因组上的比率分别为82.82%和53.15%。对于匹配率较低的情况，一般是由于5' 和3' 末端测序质量不高导致的，可以使用参数-5和-3去除5' 和3' 末端部分碱基。

步骤5：使用less SRR192340.map浏览输出定位文件的内容（图6-23）。

```
[lirh2011@webserver data]$ less SRR192340.map
SRR192340.1 HWI-EAS266:3:1:0:629 length=36      -       chr22   39032325        GCTCAGCTCAGACACCTGCAAGTCTGTGGCCCTAGG    @>;@B
A@?AA=@;>BBB?<ABAAAAB@BBBBA?BBA     0       13:C>T
SRR192340.2 HWI-EAS266:3:1:0:488 length=36      +       chr2    133012738       CCTCCGACTTTCGTTCTTGATTAATGAAAACATTCT    BBCCC
CBBCCCBCBBACCCBCBBBBCCBBCCCCBBC     0
SRR192340.3 HWI-EAS266:3:1:0:380 length=36      +       chrM    2257    CCGCACCCAATTGGACCAATCTATCACCCTATAGAA    BBA@BBA@9>?BB
B6B?><>AB@??A>@ABBB@=9A     0       0:T>C,2:A>G
SRR192340.4 HWI-EAS266:3:1:0:467 length=36      -       chrM    2229    AAAAATCCCAAACATTTAACTGAACTCCTCACACCT    =C>CCCCCCCCC
CCCCCCCCCCCCCCCCCBCAC;?     0       0:C>T,20:A>T
SRR192340.1 HWI-EAS266:3:1:0:629 length=36      -       chr22   39032325        GCTCAGCTCAGACACCTGCAAGTCTGTGGCCCTAGG    @>;@B
AAAB@BBBBA?BBA     0       13:C>T
SRR192340.2 HWI-EAS266:3:1:0:488 length=36      +       chr2    133012738       CCTCCGACTTTCGTTCTTGATTAATGAAAACATTCT    BBCCC
SRR192340.1 HWI-EAS266:3:1:0:629 length=36      -       chr22   39032325        GCTCAGCTCAGACACCTGCAAGTCTGTGGCCCTAGG    @>;@B
A@?AA=@;>BBB?<ABAAAAB@BBBBA?BBA     0       13:C>T
SRR192340.2 HWI-EAS266:3:1:0:488 length=36      +       chr2    133012738       CCTCCGACTTTCGTTCTTGATTAATGAAAACATTCT    BBCCC
CBBCCCBCBBACCCBCBBBBCCBBCCCCBBC     0
SRR192340.3 HWI-EAS266:3:1:0:380 length=36      +       chrM    2257    CCGCACCCAATTGGACCAATCTATCACCCTATAGAA    BBA@BBA@9>?BB
B6B?><>AB@??A>@ABBB@=9A     0       0:T>C,2:A>G
SRR192340.4 HWI-EAS266:3:1:0:467 length=36      -       chrM    2229    AAAAATCCCAAACATTTAACTGAACTCCTCACACCT    =C>CCCCCCCCC
CCCCCCCCCCCCCCCCCBCAC;?     0       0:C>T,20:A>T
```

图6-23 bowtie输出.map文件的内容

文件第1列是ID和长度信息，第2列表示匹配在正链（+）还是负链（-）上，第3列表示匹配的染色体，第4列表示匹配的起始位置，第5列是Read序列，第6列是质量数，第7列表示有多少条序列以相同的错误率匹配在该位置，0表示这条Read是该位置的最佳匹配。关于该文件格式的详细描述请参考：http: //bowtie-bio.sourceforge.net/manual.shtml#default-bowtie-output。

步骤6：bowtie还可以产生SAM格式的输出文件（图6-24），适用于其他支持SAM格式文件的软件进一步分析。对于单末端测序数据SRR192340，执行命令bowtie –t –S --sam-nohead --sam-nosq hg19 SRR192340.fastq SRR192340.sam，其中参数-S/--sam 代表以sam格式输出定位结果，在后面加入--sam-nohead来去掉所有的SAM表头，--sam-nosq来去掉所有@SQ的表头（例如某个序列比对到了参考基因组的许多位置）。

对于双末端测序数据SRR192333，执行命令bowtie –t –p 8 –I 200 –X 1000–S --sam-

```
[lirh2011@webserver data2]$ head -400000 /pub2/lirh2011/book/data/SRR192333_1.fastq > /pub2/lirh2011/book/data2/SRR192333_1.f
astq
[lirh2011@webserver data2]$ head -400000 /pub2/lirh2011/book/data/SRR192333_2.fastq > /pub2/lirh2011/book/data2/SRR192333_2.f
astq
[lirh2011@webserver data2]$ head -400000 /pub2/lirh2011/book/data/SRR192340.fastq > /pub2/lirh2011/book/data2/SRR192340.fastq
[lirh2011@webserver data2]$ bowtie -t -p 8 -I 200 -X 1000 -S --sam-nohead --sam-nosq --chunkmbs 640 /pub2/lirh2011/book/my_in
dexes/hg19 -1 /pub2/lirh2011/book/data2/SRR192333_1.fastq -2 /pub2/lirh2011/book/data2/SRR192333_2.fastq /pub2/lirh2011/book/
data2/SRR192333.sam
Time loading reference: 00:00:25
Time loading forward index: 00:00:06
Time loading mirror index: 00:00:16
Seeded quality full-index search: 00:00:13
# reads processed: 100000
# reads with at least one reported alignment: 47707 (47.71%)
# reads that failed to align: 52293 (52.29%)
Reported 47707 paired-end alignments to 1 output stream(s)
Time searching: 00:01:01
Overall time: 00:01:01
[lirh2011@webserver data2]$ bowtie -t  -S --sam-nohead --sam-nosq  /pub2/lirh2011/book/my_indexes/hg19 /pub2/lirh2011/book/da
ta2/SRR192340.fastq  /pub2/lirh2011/book/data2/SRR192340.sam
Time loading forward index: 00:00:02
Time loading mirror index: 00:00:01
Seeded quality full-index search: 00:00:10
# reads processed: 100000
# reads with at least one reported alignment: 82582 (82.58%)
# reads that failed to align: 17418 (17.42%)
Reported 82582 alignments to 1 output stream(s)
Time searching: 00:00:14
Overall time: 00:00:14
```

图6-24 bowtie输出.sam文件

nohead --sam-nosq hg19 -1 SRR192333_1.fastq -2 SRR192333_2.fastq SRR192333.sam，其中参数-I表示paired end所测两序列的最短距离（含两序列长度），本例中，双侧序列都为75bp，内部空缺序列为50bp，因此选200；-X指定paired end所测两序列的最长距离，考虑到可变剪切跨越外显子，本例中我们选择1000。当使用多核CPU时，可以使用-p参数指定使用多少线程计算。为缩短运行时间，我们只选取fastq文件的部分read进行参考基因组定位。

输出的sam格式文件如下（图6-25）：

```
[lirh2011@webserver data]$ less SRR192333.sam
SRR192333.1 HWUSI-EAS1671_0001:5:1:1022:10290 length=75 77      *       0       0       *       *       0       0       NCCAC
TTCTACGATGCCATGGATGGGCAGATACAGGGCAGCGTGGAGCAGGCAGGCCAAGGNNNGGNNAAGGAAG      ##################################################
#########################       XM:i:0
SRR192333.1 HWUSI-EAS1671_0001:5:1:1022:10290 length=75 141     *       0       0       *       *       0       0       NGCAG
CTNNNNNAGGACTNNNNNNNNNNGCTNNNNCNNGNNCCTCCTCTGGAGCAAAACATGACCGGCGTTGGGG      ##################################################
#########################       XM:i:0
SRR192333.2 HWUSI-EAS1671_0001:5:1:1022:15574 length=75 77      *       0       0       *       *       0       0       NCAAG
CAGGCCCCCACCTGCCCTTAGTGATGTTTGGAGTCGTTTTACCCTCTTCTATTGAANNNCCNNGGGATTT      ##################################################
#########################       XM:i:0
SRR192333.2 HWUSI-EAS1671_0001:5:1:1022:15574 length=75 141     *       0       0       *       *       0       0       NTGCT
TCANNNNTTTTTCNNNNNNNNNTNTGCNNNNGNNGNNAGGCAGCCTTCTACCATTTCCTCAGCAACAGAGG      ##################################################
#########################       XM:i:0
SRR192333.3 HWUSI-EAS1671_0001:5:1:1022:17698 length=75 77      *       0       0       *       *       0       0       NAAGG
GGTACAGGGCTGGAGCCAAGGCACTTCTGGGCAGCTTCATCAGCTACTGTTCTTGGNNNAACNTCCTGAG      #+++++*))):76::AAAAAAAAAAAA87AAAAAA7AAAAAAAAA78A7A
AAAAAAAAA################       XM:i:0
SRR192333.3 HWUSI-EAS1671_0001:5:1:1022:17698 length=75 141     *       0       0       *       *       0       0       NAGGA
AGTNNNNCCAAGANNNNNNNNTNATGNNNNTNNCNNGAAGTGCCTTGGCTCCAGCCCTGTACCCCTTGAG      ##################################################
#########################       XM:i:0
SRR192333.4 HWUSI-EAS1671_0001:5:1:1022:4778 length=75  77      *       0       0       *       *       0       0       NTTGA
TTCAAGTTCCTTGGATCCTATCCTGACTTGGACATTTAGGCAAGAGCTAACTTGGGNNNAGNNTTAACTT      #((('*+***AA666,*3,3333,375868AAAA65553-A56A0A8AA6
AA55A####################       XM:i:0
SRR192333.4 HWUSI-EAS1671_0001:5:1:1022:4778 length=75  141     *       0       0       *       *       0       0       NGGAC
AGNNNNNCCTGTGNNNNNNNNGNTACNNNNGNNANNGTGGAGAATCAGACCTGAGAGGTGGAGGCAGAGG      ##################################################
#########################       XM:i:0
```

图6-25　bowtie输出.sam文件的内容

其中第1列是ID，第2列是该序列满足各标签代表数字的和，具体如下：

数字	标签
1	read是一个pair中的一个
2	这个比对是一个合适的paired end比对的一个末端
4	这个read没有匹配上
8	read是一个pair中的一个，并且没有匹配上
16	匹配到负链上了
32	在paired end比对中的另一条序列匹配到参考基因组的负链上
64	在一个pairedread第一个（#1）
128	在一个pairedread第二个（#2）

其他列代表含义请参考

http://bowtie-bio.sourceforge.net/manual.shtml#default-bowtie-output。

四、短序列在参考基因组上定位和可变剪切的识别

软件TopHat不仅能够将短序列定位到参考基因组，还能够识别转录本序列剪切、插入和删除等信息，以下分六个步骤介绍TopHat软件的安装、应用以及结果的可视化。

步骤1：访问TopHat软件网站http://tophat.cbcb.umd.edu/index.html（图6-26），下载软件。

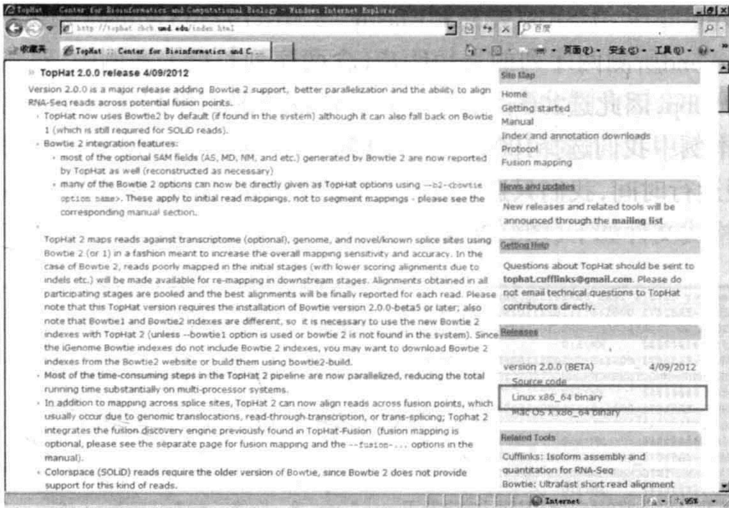


图6-26 软件tophat首页

步骤2：安装tophat。将下载的tophat-2.0.0.Linux_x86_64.tar.gz文件上传到服务器，使用tar命令将文件解压缩，然后将得到的目录tophat-2.0.0.Linux_x86_64添加到linux系统的环境变量PATH中（在文件~/.bash_profile文件中添加export PATH= / tophat-2.0.0.Linux_x86_64：$PATH）。

步骤3：准备参考基因组、参考基因组注释文件、测序数据fastq格式文件，并上传到服务器。参考基因组文件可以从cufflinks网站下载http: //cufflinks.cbcb.umd.edu/igenomes.html; 为了计算某个基因或某个转录本等所映射的read读数，可以从Ensembl数据库中下载基因组注释文件ftp: //ftp.ensembl.org/pub/release-66/gtf/homo_sapiens/Homo_sapiens.GRCh37.66.gtf.gz。

步骤4：执行命令定位Read。对于双末端测序数据SRR192333命令如下：tophat -p 8 -r 50 -G Homo_sapiens.NCBI36.54.gtf -o SRR192333 hg19 SRR192333_1.fastq SRR192333_2.fastq，其中，-r 50表示双末端序列间距50个碱基；-G参数指定基因组注释文件（图6-27）。



图6-27 执行tophat定位短序列

对于单末端测序数据SRR192340命令如下：tophat –p 8　–G Homo_sapiens.GRCh37.66.gtf –o hg19 SRR192340.fastq。

步骤5：查看结果。tophat输出的主要结果包括reads在参考基因组上的匹配列表accepted_hits.bam，剪切方式文件junctions.bed，插入和删除信息deletions.bed和insertions.bed。其中，bed文件可以通过UCSC Genome Browser(http: //genome.ucsc.edu/)可视化。

步骤6：可视化剪切方式文件junctions.bed。有关bed格式文件的介绍请访问UCSC。进入主页http: //genome.ucsc.edu/后，点击Genome Browser进入，在assembly的下拉列表中，选择分析所用到的参考基因组(图6–28)，此处选择hg19。

点击add custom tracks打开自定义轨道管理页面(图6–29)。

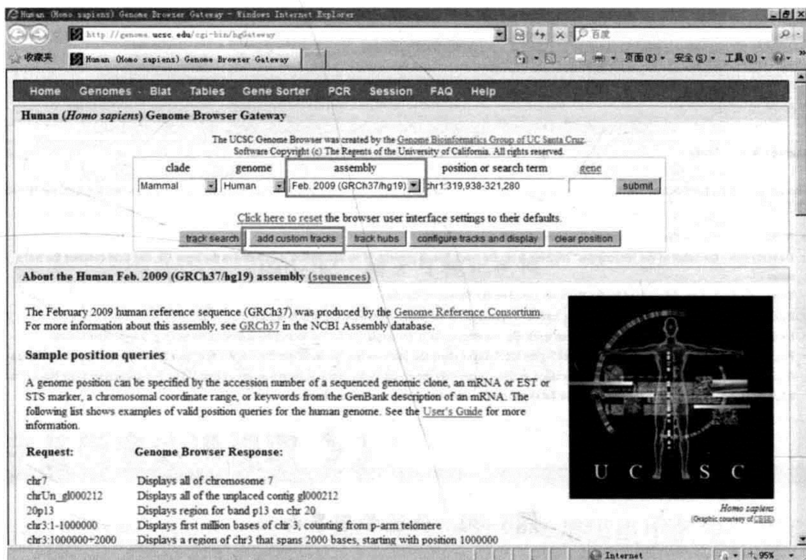


图6–28　UCSC数据库Genome Browser页面

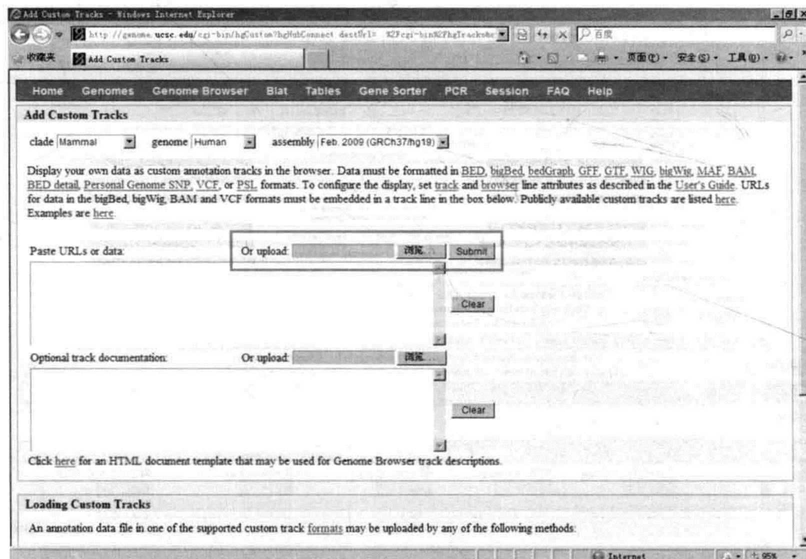


图6–29　UCSC中添加自定义轨道

点击浏览选择tophat输出的junctions.bed文件，按submit提交，会显示在轨道管理页面（图6-30）。

点击go to genome browser回到Genome Browser页面。在基因组图的上方出现TopHat junctions的图示，页面下方自定义轨道Custom Tracks里面有junctions选项，选择full，Genes and Gene Prediction Tracks里面的Human ESTs，Spliced ESTs都选择dense，点击右侧的refresh（图6-31）。

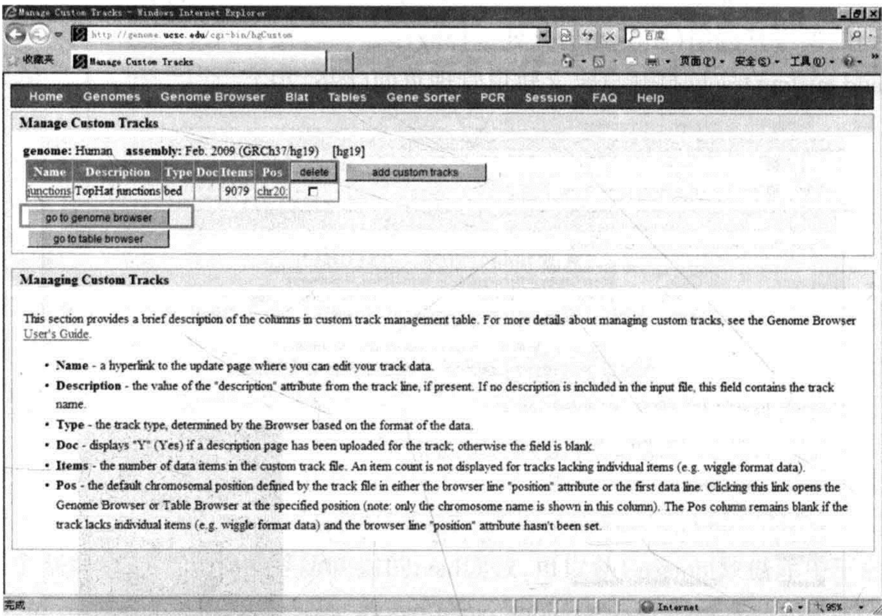


图6-30 轨道管理界面

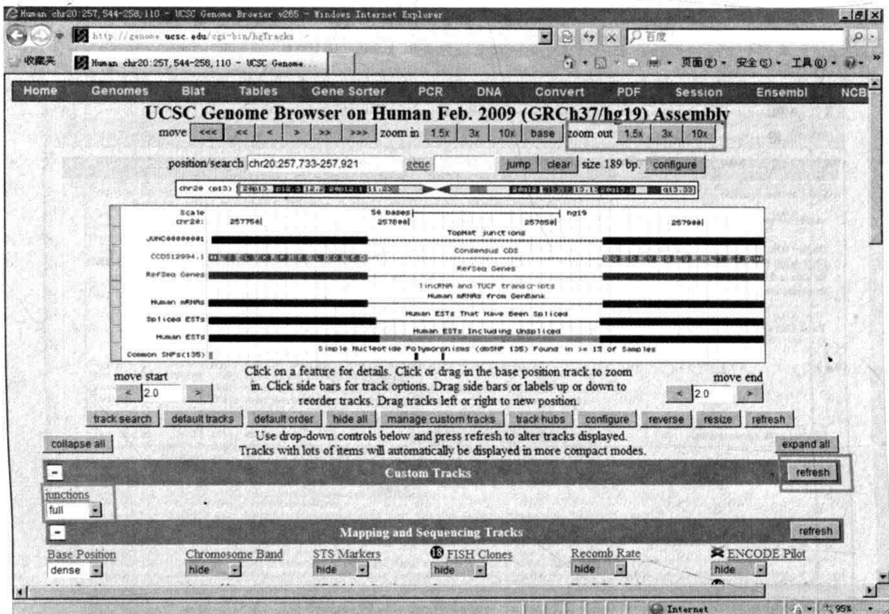


图6-31 可视化剪切方式文件junctions.bed

可以通过上方的工具条move，zoom in以及zoom out可以移动、放大和缩小图示，方便浏览。例如，点击两次页面上方右侧zoom out 10x按钮，使基因组的显示范围增大100倍，如下（图6-32）。

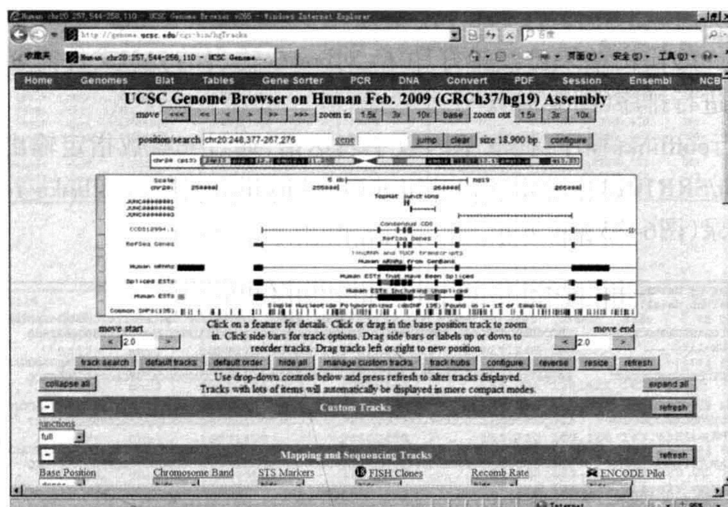


图6-32　增大可视化范围

页面显示了20号染色体第248 377个碱基到第267 276个碱基范围的基因组图。其中包含3个由TopHat软件估计出的剪切方式，与已知的剪切方式相同。

五、数字基因表达谱提取

软件cufflinks能够根据tophat软件得到的短序列在参考基因组中的定位信息以及基因在参考基因组中的注释文件，计算出基因的数字表达谱，同时还能计算不同样本间基因、转录本等的差异表达，识别选择性剪切等。以下分七个步骤介绍cufflinks的下载、安装及应用。

步骤1：访问Cufflinks软件网站http: //cufflinks.cbcb.umd.edu/，下载软件（图6-33）。

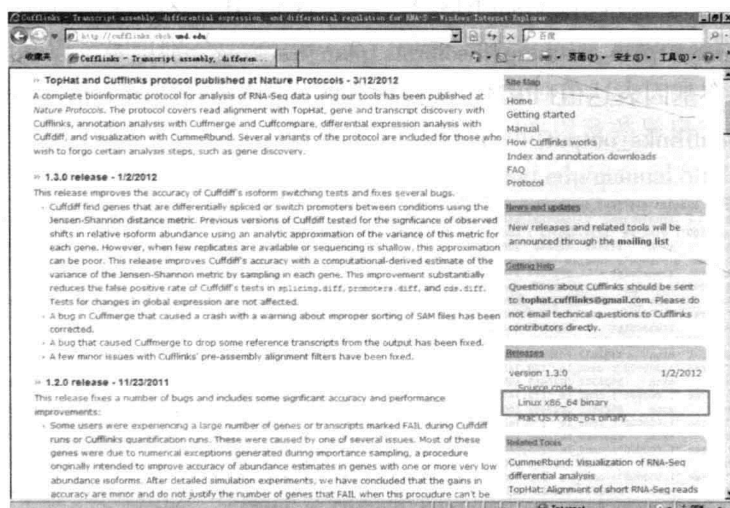


图6-33　Cufflinks软件下载

步骤2：安装cufflinks。将下载的cufflinks-1.3.0.Linux_x86_64.tar.gz文件上传到服务器，使用tar命令将文件解压缩，然后将得到的目录cufflinks-1.1.0.Linux_x86_64添加到linux系统的环境变量PATH中（在文件~/.bash_profile文件中添加export PATH= / cufflinks-1.1.0.Linux_x86_64：$PATH）。

步骤3：tophat的输出结果已经排好序了，对bowtie输出的SAM文件需要使用sort命令根据染色体和位置进行排序（图6-34）。

步骤4：运行cufflinks程序，提取基因数字表达谱，通过-o参数指定输出目录。cufflinks -o cufflinks_result/SRR192333 SRR192333/accepted_hits.bam，其中cufflinks_result/SRR192333是指定的输出目录（图6-35）。

```
[lirh2011@webserver data2]$ sort -k 3,3 -k 4,4n SRR192333.sam>SRR192333.sam.sorted
[lirh2011@webserver data2]$ less SRR192333.sam.sorted
SRR192333.85821 99      chr1    14580   255     75M     =       14729   224     CGCTGGTTCCGTCACCCCCTCCCAAGGAAGTAGGTCTGAGCAGCT
TGTCCTGGCTGTGTCCATGTCAGAGCAACG      CCCCCCCCCCCCCCCCCCBBBBBBCCCCCBCCCACCCCCCCCCCCADCCCCBBCC?@?BCBCBBDDD:BAB@BB      XA:i:1  MD
:Z:0T74         NM:i:1
SRR192333.85821 147     chr1    14729   255     75M     =       14580   -224    CTGTGGCTGCTGCGGTGGCGGCAGAGGAGGGATGGAGTCTGACAC
GCGGGCAAAGGCTCCTCCGGGCCCCTCACC      ##?@A75077(2(C>ABBBBA=8@3BABBBD>>AB@@B?A2B@BABBDCCCCCDCACACD@CACCCCCCCCCBCC      XA:i:0  MD
:Z:75 NM:i:0
SRR192333.97584 99      chr1    461376  255     75M     =       461507  206     CCGTGCTTTTCCCAAAGGTTGTCTGGGGACCTCAGTAAGTAAAGG
GGAGAAGTGTGGGTGTTGGGGAAAGGGGAA      CCCCCCCCCCCCCCCCCCDCCCCCCCCCCCDCCCCDCCCBCDDCCCDBCB@CBCDCAAACBABBBDAA?DBBB??      XA:i:1  MD
:Z:17A57        NM:i:1
SRR192333.97584 147     chr1    461507  255     75M     =       461376  -206    TTTGGTCAGCGTTCACTGAATACACATTTTACATGTGATGGAGGT
AGAGGCAGGGATGTAGCTTTTTTATCTTTG      BDBDDBDBCDBCBCBBCCCCC>CDCCCCCCDCCCCCCCCCCCCCCCCCCCCCCCCCCCBBBCCCCCCCCCCC      XA:i:0  MD
:Z:75 NM:i:0
SRR192333.17593 99      chr1    564464  255     75M     =       564603  214     GGGAGTCCGAACTAGTCTCAGGCTTCAACATCGAATACGCCGCAG
GCCCCTTCGCCCTATTCTTCATAGCCGAAT      CCCCCCCCCCCCCDCCCCCCCACCCCCCCCCCCAACCCBCCBCCDACCCC@BCBBCB=>>@B=BCB>>8CCB?.?      XA:i:0  MD
:Z:75 NM:i:0
```

图6-34 排序bowtie输出的sam文件

```
[lirh2011@webserver data2]$ cufflinks -o cufflinks_result/SRR192333 SRR192333/accepted_hits.bam
cufflinks: /usr/lib64/libz.so.1: no version information available (required by cufflinks)
[15:48:05] Inspecting reads and determining fragment length distribution.
> Processed 38711 loci.                        [*************************] 100%
> Map Properties:
>       Total Map Mass: 87413.95
>       Read Type: 72bp paired-end
>       Fragment Length Distribution: Gaussian (default)
>                     Estimated Mean: 208.93
>                  Estimated Std Dev: 70.97
[15:48:21] Assembling transcripts and estimating abundances.
> Processed 38711 loci.                        [*************************] 100%
```

图6-35 运行cufflinks提取基因表达谱

步骤5：cufflinks输出3个文件，提供了不同水平的表达transcripts.gtf（图6-36）、isoforms.fpkm_tracking和 genes.fpkm_tracking（图6-37、图6-38）。其中transcripts.gtf记录的是cufflinks装配的异构体，genes.fpkm_tracking和isoforms.fpkm_tracking分别记录了以基因为单位和以转录本为单位的数字基因表达值FPKM。文件内容的详细说明请参考http: //cufflinks.cbcb.umd.edu/manual.html#cufflinks_output。

```
[lirh2011@webserver SRR192333]$ less transcripts.gtf
chr1    Cufflinks       transcript      568468  568845  1000    .       .       gene_id "CUFF.47"; transcript_id "CUFF.47.1"; FPKM "1303.94880
72042"; frac "1.000000"; conf_lo "1231.728345"; conf_hi "1376.169270"; cov "16.422596";
chr1    Cufflinks       exon    568468  568845  1000    .       .       gene_id "CUFF.47"; transcript_id "CUFF.47.1"; exon_number "1"; FPKM "1
303.9488072042"; frac "1.000000"; conf_lo "1231.728345"; conf_hi "1376.169270"; cov "16.422596";
chr1    Cufflinks       transcript      1717098 1718264 1000    .       .       gene_id "CUFF.295"; transcript_id "CUFF.295.1"; FPKM "135.6913
207871"; frac "1.000000"; conf_lo "112.393997"; conf_hi "158.988644"; cov "1.710766";
chr1    Cufflinks       exon    1717098 1718264 1000    .       .       gene_id "CUFF.295"; transcript_id "CUFF.295.1"; exon_number "1"; FPKM
"135.6913207871"; frac "1.000000"; conf_lo "112.393997"; conf_hi "158.988644"; cov "1.710766";
chr1    Cufflinks       transcript      8022872 8045216 1000    +       .       gene_id "CUFF.587"; transcript_id "CUFF.587.1"; FPKM "198.9534
282041"; frac "1.000000"; conf_lo "170.743258"; conf_hi "227.163599"; cov "2.608696";
chr1    Cufflinks       exon    8022872 8022935 1000    +       .       gene_id "CUFF.587"; transcript_id "CUFF.587.1"; exon_number "1"; FPKM
"198.9534282041"; frac "1.000000"; conf_lo "170.743258"; conf_hi "227.163599"; cov "2.608696";
chr1    Cufflinks       exon    8025384 8025485 1000    +       .       gene_id "CUFF.587"; transcript_id "CUFF.587.1"; exon_number "2"; FPKM
"198.9534282041"; frac "1.000000"; conf_lo "170.743258"; conf_hi "227.163599"; cov "2.608696";
chr1    Cufflinks       exon    8029405 8029464 1000    +       .       gene_id "CUFF.587"; transcript_id "CUFF.587.1"; exon_number "3"; FPKM
"198.9534282041"; frac "1.000000"; conf_lo "170.743258"; conf_hi "227.163599"; cov "2.608696";
chr1    Cufflinks       exon    8030954 8031023 1000    +       .       gene_id "CUFF.587"; transcript_id "CUFF.587.1"; exon_number "4"; FPKM
"198.9534282041"; frac "1.000000"; conf_lo "170.743258"; conf_hi "227.163599"; cov "2.608696";
chr1    Cufflinks       exon    8037712 8037798 1000    +       .       gene_id "CUFF.587"; transcript_id "CUFF.587.1"; exon_number "5"; FPKM
"198.9534282041"; frac "1.000000"; conf_lo "170.743258"; conf_hi "227.163599"; cov "2.608696";
chr1    Cufflinks       exon    8044954 8045216 1000    +       .       gene_id "CUFF.587"; transcript_id "CUFF.587.1"; exon_number "6"; FPKM
"198.9534282041"; frac "1.000000"; conf_lo "170.743258"; conf_hi "227.163599"; cov "2.608696";
```

图6-36 cufflinks输出的transcripts.gtf

trans_id	bundle_id	chr	left	right	FPKM	FMI	frac	FPKM_conf_lo	FPKM_conf_hi	coverage	length	effective_length	status
CUFF.47.1	38735	chr1	568467	568845	1303.95	1	1	1231.73	1376.17	16.4226	378	307	OK
CUFF.295.1	38845	chr1	1717097	1718264	135.691	1	1	112.394	158.989	1.71077	1167	1096	OK
CUFF.587.1	38970	chr1	8022871	8045216	198.953	1	1	170.743	227.164	2.6087	646	575	OK
CUFF.649.1	38999	chr1	9789110	9790022	190.437	1	1	162.837	218.037	2.40785	912	841	OK
CUFF.1297.1	39272	chr1	20520705	20521599	139.001	1	1	115.422	162.581	1.64034	894	823	OK
CUFF.1449.1	39335	chr1	22418532	22419430	179.828	1	1	153.008	206.648	2.26723	898	827	OK
CUFF.1517.1	39364	chr1	24019108	24022902	1333.48	1	1	1260.44	1406.51	15.8266	560	489	OK
CUFF.1775.1	39471	chr1	26801673	26802409	143.112	1	1	119.186	167.038	1.80948	736	665	OK
CUFF.2111.1	39620	chr1	32508152	32509440	112.8	1	1	91.5587	134.042	1.47905	1288	1217	OK
CUFF.2203.1	39658	chr1	33238489	33239830	180.155	1	1	153.31	206.999	2.18504	1341	1270	OK
CUFF.2397.1	39742	chr1	36552558	36553834	187.538	1	1	160.149	214.927	2.23547	681	610	OK
CUFF.2493.1	39781	chr1	38023237	38030745	203.676	1	1	175.133	232.219	2.15704	745	674	OK
CUFF.2775.1	39904	chr1	43391912	43392783	142.998	1	1	119.081	166.914	1.6875	871	800	OK
CUFF.2877.1	39942	chr1	44088162	44089220	185.448	1	1	158.212	212.684	2.05167	1058	987	OK
CUFF.2967.1	39979	chr1	45241715	45243800	557.393	1	1	510.174	604.611	6.61253	502	431	OK
CUFF.3015.1	39998	chr1	45976761	45977085	542.6	1	1	496.013	589.188	6.64032	324	253	OK
CUFF.3019.1	39998	chr1	45980564	45987546	819.471	1	1	762.219	876.724	9.75008	420	349	OK
CUFF.3037.1	40004	chr1	46085719	46087106	262.71	1	1	230.293	295.127	3.2881	1029	958	OK
CUFF.3093.1	40029	chr1	46646211	46651628	1699.36	1	1	1616.91	1781.81	20.5394	1276	1205	OK
CUFF.3155.1	40056	chr1	47264746	47279685	328.83	1	1	292.563	365.097	3.85779	732	661	OK
CUFF.3157.1	40057	chr1	47279896	47284135	369.487	1	1	331.043	407.931	4.09297	814	743	OK
CUFF.3159.1	40057	chr1	47284301	47285010	430.338	1	1	388.849	471.827	5.13883	709	638	OK

图6-37　**isoforms.fpkm_tracking文件内容**

gene_id	bundle_id	chr	left	right	FPKM	FPKM_conf_lo	FPKM_conf_hi	status
CUFF.47	38735	chr1	568467	568845	1303.95	1231.73	1376.17	OK
CUFF.295	38845	chr1	1717097	1718264	135.691	112.394	158.989	OK
CUFF.587	38970	chr1	8022871	8045216	198.953	170.743	227.164	OK
CUFF.649	38999	chr1	9789110	9790022	190.437	162.837	218.037	OK
CUFF.1297	39272	chr1	20520705	20521599	139.001	115.422	162.581	OK
CUFF.1449	39335	chr1	22418532	22419430	179.828	153.008	206.648	OK
CUFF.1517	39364	chr1	24019108	24022902	1333.48	1260.44	1406.51	OK
CUFF.1775	39471	chr1	26801673	26802409	143.112	119.186	167.038	OK
CUFF.2111	39620	chr1	32508152	32509440	112.8	91.5587	134.042	OK
CUFF.2203	39658	chr1	33238489	33239830	180.155	153.31	206.999	OK
CUFF.2397	39742	chr1	36552558	36553834	187.538	160.149	214.927	OK
CUFF.2493	39781	chr1	38023237	38030745	203.676	175.133	232.219	OK
CUFF.2775	39904	chr1	43391912	43392783	142.998	119.081	166.914	OK
CUFF.2877	39942	chr1	44088162	44089220	185.448	158.212	212.684	OK
CUFF.2967	39979	chr1	45241715	45243800	557.393	510.174	604.611	OK
CUFF.3015	39998	chr1	45976761	45977085	542.6	496.013	589.188	OK
CUFF.3019	39998	chr1	45980564	45987546	819.471	762.219	876.724	OK
CUFF.3037	40004	chr1	46085719	46087106	262.71	230.293	295.127	OK
CUFF.3093	40029	chr1	46646211	46651628	1699.36	1616.91	1781.81	OK
CUFF.3155	40056	chr1	47264746	47279685	328.83	292.563	365.097	OK

图6-38　**genes.fpkm_tracking文件内容**

步骤6：重复步骤1到4，处理测序数据SRR192336。

步骤7：简单的差异基因筛选。使用cufflinks软件的cuffdiff命令能够计算差异表达的基因、转录本、选择性剪切、启动子，以健康样本GSM718707的数据SRR192333，与肺癌样本GSM718710的数据SRR192336 为例，基于tophat得到的结果，运行cuffdiff命令cuffdiff Homo_sapiens.GRCh37.66.gtf SRR192333/accepted_hits.bam SRR192336/accepted_hits.bam运行结果分为四种类型：四个水平的表达数据，转录本表达isoforms.fpkm_tracking、基因表达genes.fpkm_tracking 、编码序列表达cds.fpkm_tracking以及初级转录本表达tss_groups.fpkm_tracking；对应四个水平的差异表达检验数据isoforms_exp.diff、genes_exp.diff 、cds_exp.diff以及tss_groups_exp.diff；一个差异剪切数据splicing.diff;差异编码数据cds.diff;以及差异启动子数据promoters.diff。每个文件内容格式详细说明见http：//cufflinks.cbcb.umd.edu/manual.html#cuffdiff_output 。这四个水平的差异表达数据集就是我们需要的差异表达基因或差异剪切，根据需要可以解下再进行功能分析等。

（张绍军　李茉红　宫滨生　李　霞）

参考文献

1. Ansorge WJ. Next-generation DNA sequencing techniques. N Biotechnol 2009，25(4)：195-203.
2. Erdmann J. Next generation technology edges genome sequencing toward the clinic. Chem Biol，2011，18(12)：

1513–1514.

3. Ingolia NT. Genome–wide translational profiling by ribosome footprinting. Methods Enzymol 2010,470 : 119–142.
4. Mardis ER. The impact of next–generation sequencing technology on genetics. Trends Genet 2008,24(3): 133–141.
5. Mardis ER. Next–generation DNA sequencing methods. Annu Rev Genomics Hum Genet 2008,9 : 387–402.
6. Nowrousian M. Next–generation sequencing techniques for eukaryotic microorganisms: sequencing–based solutions to biological problems. Eukaryot Cell,2010 ,9(9): 1300–1310.
7. Rothberg JM, Leamon JH. The development and impact of 454 sequencing. Nat Biotechnol 2008,26(10): 1117–1124.
8. Shendure J, Ji H. Next–generation DNA sequencing. Nat Biotechnol 2008,26(10): 1135–1145.
9. Trapnell C, Salzberg SL. How to map billions of short reads onto genomes. Nat Biotechnol 2009,27(5): 455–457.
10. Zhou X, Ren L, Meng Q, et al. The next–generation sequencing technology and application. Protein Cell,2010, 1(6): 520–536.

第七章

CHAPTER 7

转录调控的信息学分析

BIOINFORMATICS ANALYSIS ON TRANSCRIPTION REGULATION

基因表达是指细胞在生命过程中,把存储在DNA中的遗传信息经过转录、剪接、翻译以及翻译后修饰等过程,转变为具有生物活性的蛋白质分子。作为基因表达过程的第一步,转录在基因表达过程中起到了至关重要的作用。哺乳动物有机体约含30 000~40 000个基因,它们是如何在适当的空间和时间进行转录调控的,转录因子(TFs)是如何在这些差异表达的基因间行使功能的,我们需要明确转录调控机制,开发信息学方法来解决这些问题。本章将介绍如何应用信息学方法研究在转录过程中起到关键作用的启动子、转录因子、可变剪接等功能位点的特征,探讨转录调控机制与人类疾病之间的关联,以及如何应用信息学方法综合分析这些关联。

第一节

基因的转录调控

Section 1 Transcription Regulation

一、转录

转录是以一条DNA链为模板，通过酶的激活作用，利用碱基互补配对的原则合成一条与模板DNA反向平行且互补的RNA的过程。其中，作为模板的DNA链被称为"模板链"、"负链"或者"反义链"（图7-1）。转录与DNA自我复制结果的不同在于，在转录产物RNA里，尿嘧啶（U，uracil）替代了DNA复制结果中的胞嘧啶（T，thymine）。

5' ACATCGACGCGCAGTTAATCCC...3' DNA编码链（+）
3' TGTAGCTGCGCGTCAATTAGGG...5' DNA模板链（-）

5' ACAUCGACGCGCAGUUAAUCCC...3' 产物RNA链（+）

图7-1 转录过程中的模板DNA链与产物RNA链

一个完整的转录过程可以总结成以下四个中心步骤：

1. 聚合酶结合到转录起始位点 DNA序列上起始转录的信号称为启动子。原核生物聚合酶可以识别启动子并直接与之结合。而真核生物聚合酶需要借助其他蛋白质，这些蛋白质被称为转录因子。

2. 解开DNA双螺旋（图7-2） 能够解开DNA双螺旋结构的酶称为解旋酶。原核生物聚合酶具有解旋活性，而真核生物聚合酶没有解旋活性。所以真核生物DNA双螺旋的解开需要借助于一类特殊的转录因子。

3. 基于DNA模板链合成RNA RNA聚合酶使用三磷酸核苷（NTPs）构造一条RNA链。

4. 合成的终止 原核生物和真核生物使用不同的信号来终止转录。

在真核细胞中，新合成的RNA链在完成3'加poly-A尾和5'加帽后，通过核孔复合物出细胞核，进入到细胞质中。真核生物的转录过程比原核生物更复杂。一个方面的原因是真核生物DNA被组蛋白缠绕，可以阻止聚合酶接近启动子。

转录是基因表达的第一步。由DNA转录成的一个RNA分子称为一个转录单元，它可以编码至少一个基因。如果这个基因是编码蛋白质的基因，那么转录成的RNA被称为信使RNA（mRNA）。另外，转录出的基因还可能编码核糖体RNA（rRNA）、转运RNA（tRNA）、蛋白装配过程中的其他组分或者核酶。

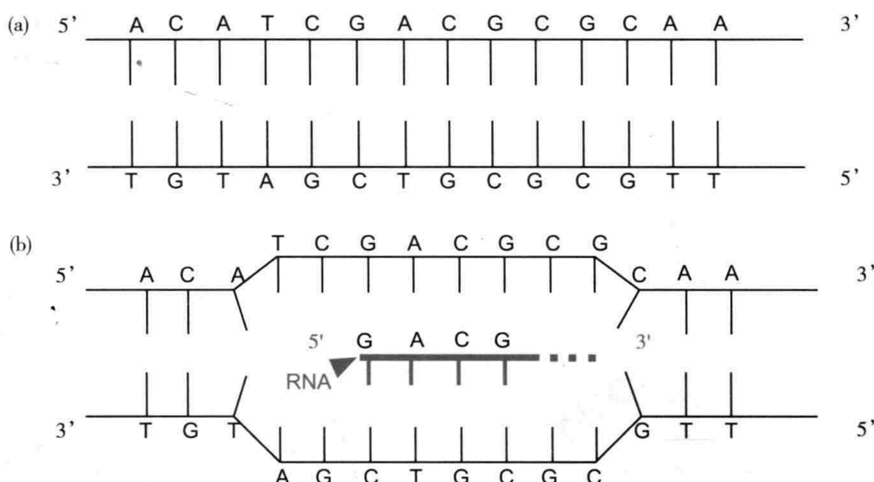


图7-2　转录过程中DNA双螺旋的打开

a. 转录前的DNA; b. 在转录过程中，DNA必须经过解旋，使得其中一条链能够作为模板合成互补RNA

一个DNA转录单元并不是全被翻译成蛋白质序列(编码序列)，其中还包含调控蛋白质合成的调控序列。编码序列上游的调控序列称为5' 非翻译区(5'UTR)，编码序列下游的调控序列称为3' 非翻译区(3'UTR)。

二、RNA聚合酶

RNA聚合酶是一类指导合成RNA的酶。在细胞中，RNA聚合酶以DNA基因作为模板合成RNA链，完成基因的转录。RNA聚合酶存在于所有有机体以及许多病毒中，它最早是由Sam Weiss、Audrey Stevens以及Jerard Hurwitz在1960年独立发现。2006年的诺贝尔化学奖被授予Roger Kornberg，以表彰他在描绘转录的不同阶段中RNA聚合酶的分子影响研究中所做出的贡献。

(一) RNA聚合酶的分类

1. 原核生物　以线虫(E.coli)为例：

一个线虫RNA聚合酶由五个亚基组成：两个α亚基、β亚基及β'亚基各一个、以及σ亚基。其中，β(151kD)和β'(156kD)显著大于α(37kD)。目前，σ亚基的一些不同形式已经被识别出来，它们的分子质量在28kD到70kD之间。σ亚基是一个已知的σ因子，它不仅在转录起始位点的识别中发挥了重要的功能，同时控制着打开DNA双螺旋结构的解旋酶的活性。核苷酸的合成过程由其他四个亚基完成，它们被合称为核心聚合酶。“全酶”(holoenzyme)是指一个完整的并且具有全部功能的酶。在E.coli中，全酶包括核心聚合酶和σ因子(图7-3)。

2. 真核生物　根据RNA聚合酶指导合成产物的不同，真核动物的RNA聚合酶可以分成三类：分别是RNA聚合酶Ⅰ、Ⅱ和Ⅲ。每类聚合酶包含两个大亚基及12~15个小亚基。其中，两个大亚基与E.coli中的β、β'亚基同源，两个小亚基与E.coli中的α亚基相似。但是，真核

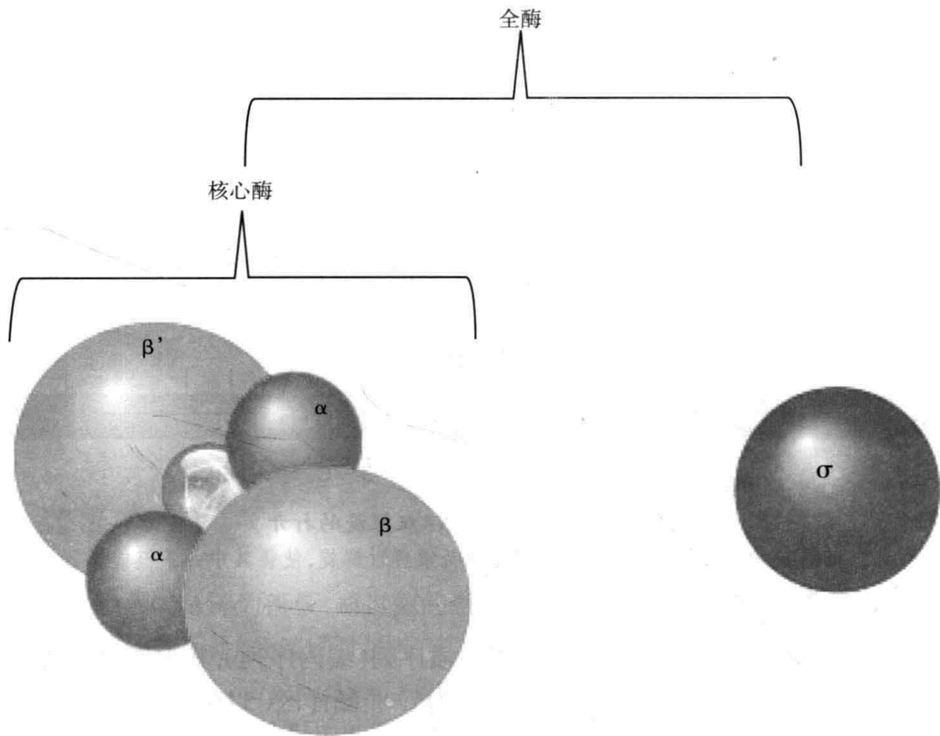


图7-3 原核生物RNA聚合酶的组成

生物RNA聚合酶中不含有任何与E.coli的σ因子相似的亚基。因此，在真核生物中，转录的起始必须由其他蛋白介导。

RNA聚合酶Ⅱ参与所有蛋白质编码基因以及大多数的snRNA基因的转录。因此，RNA聚合酶Ⅱ也成为三类RNA聚合酶中被研究最多的一类。其他两类RNA聚合酶仅仅参与RNA基因的转录。RNA聚合酶Ⅰ在核仁中，转录除5S rRNA外的所有rRNA基因。RNA聚合酶Ⅲ在核仁外，转录5S rRNA、tRNA、U6 snRNA以及一些小RNA基因。

另外，RNA聚合酶Ⅳ和Ⅴ在植物中分别指导siRNA及参与siRNA定位的异染色质形成的RNA的合成。

(二) RNA聚合酶的功能

RNA聚合酶与DNA聚合酶都具有在已存在的链上继续添加核苷酸使之延长的功能。这两类酶的主要区别在于，RNA聚合酶可以起始一条新链而DNA聚合酶并没有这个能力。因此，在DNA复制的过程中，必须先由一个不同的酶来合成一段称为引物的寡核苷酸。

三、转录调控元件

一个基因由转录区与调控区组成。转录区作为DNA的一部分被转录成初级转录本(一个与转录区DNA互补的RNA分子)。调控区可以被划分为顺式调控(cis-regulatory, or cis-acting)元件和反式调控(trans-regulatory, or trans-acting)元件。顺式调控元件是转录因子的结合位点。转录因子(一类蛋白质)与顺式调控元件结合，可以增强或抑制转录。反式调控

元件是编码转录因子的DNA序列。

顺式调控元件可以被分成以下四种类型(图7-4)。

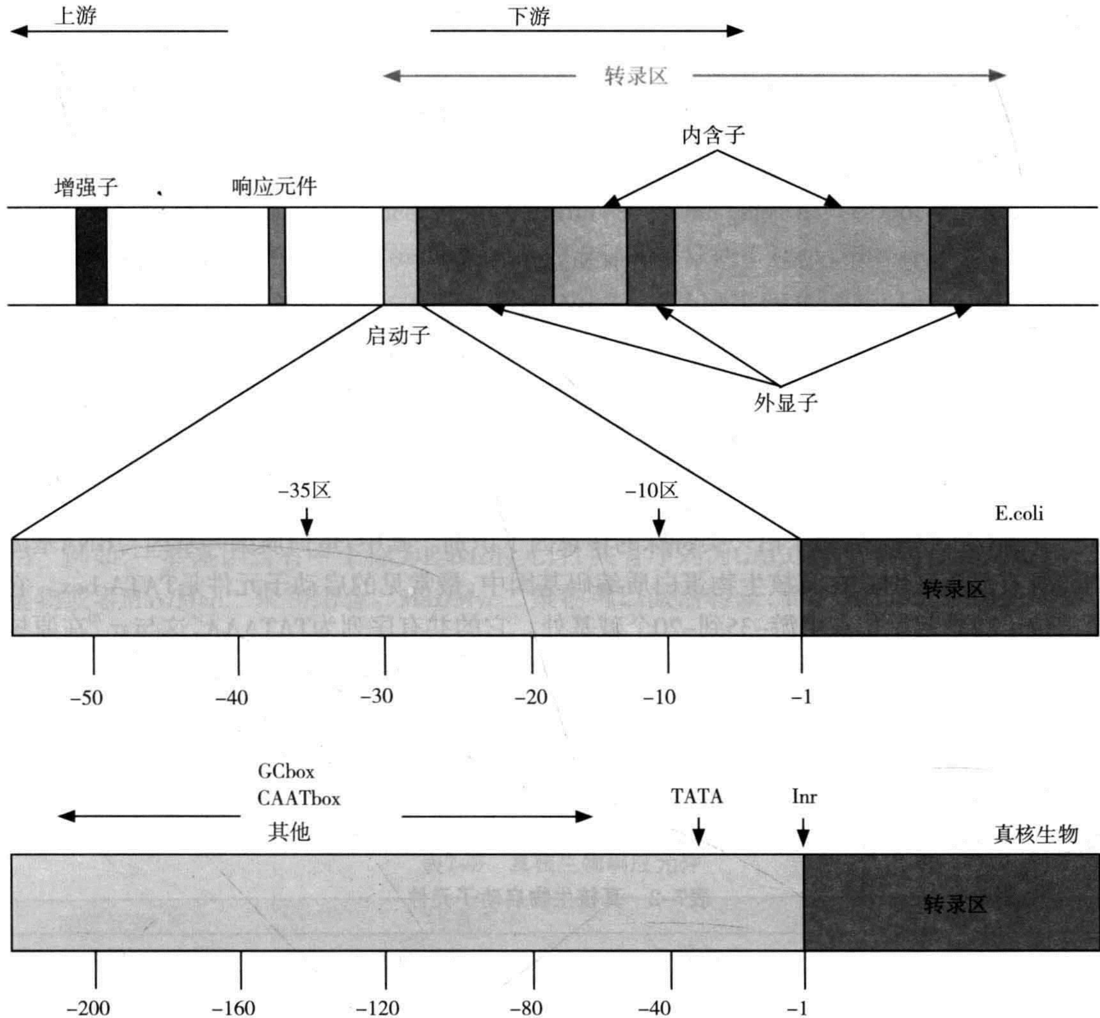


图7-4 基因的结构

转录区域包含外显子和内含子。调控元件包括启动子、应答元件、增强子和沉默子。下游(downstream)指转录进行的方向,上游(upstream)指与转录相反的方向。

(一) 启动子

启动子是DNA上转录起始的一段区域,它是一个转录开始的信息提供者,通常位于转录起始位点的上游。RNA聚合酶能够识别并与之结合,从而起始基因转录。转录的起始是基因表达的关键阶段,而这一阶段的重要问题是RNA聚合酶与启动子的相互作用。启动子的结构影响了它与RNA聚合酶的亲和力,从而影响了基因表达的水平。在原核生物中,启动子序列由RNA聚合酶中的σ因子识别。以E.coli为例, E.coli有五类σ因子: σ^{70}:调控大部分基因的表达; σ^{32}:调控热激蛋白(heat shock proteins)的表达; σ^{28}:调控鞭毛操纵子(flagellar operon)的表达(与细胞移动有关); σ^{38}:调控基因表达对抗外部压力; σ^{54}:调控

与氮代谢相关的基因表达。表7-1中概括了由E.coli的σ因子识别的启动子的共有序列(除σ^{38}外,目前并不清楚)。共有序列(consensus sequence)是与调控蛋白互作的理想序列。一个启动子通常含有与共有序列一致或者非常接近的序列元件。

表7-1 E.coli σ因子与其识别位点(启动子)的共有序列

σ因子	启动子共有序列	
	-35区	-10区
σ70	TTGACA	TATAAT
σ32	TCTCNCCCTTGAA	CCCCATNTA
σ28	CTAAA	CCGATAT
	-24区	-12区
σ54	CTGGNA	TTGCA

注:-10区又称Pribnow box,N代表任意碱基。

在真核生物中,启动子由一类特殊的转录因子识别。其中,蛋白质编码基因与RNA基因的转录有显著区别。在真核生物蛋白质编码基因中,最常见的启动子元件是TATA box。它通常位于转录起始位点上游-35到-20个碱基处。它的共有序列为TATAAA,这与σ^{70}在原核生物-10区域的识别位点极为相似。另一个启动子元件被称为起始子(initiator, Inr)。它的共有序列为PyPyAN(T/A)PyPy,其中Py代表嘧啶(C或T),N代表任意碱基,(T/A)代表T或A。在第三个位置上的碱基A位于+1,即转录起始位点。TATA box 和起始子是核心启动子元件。还有其他位于转录起始位点200bp以内的元件,如CAAT box和GC box,它们又被称为启动子邻近元件(promoter-proximal elements)。真核生物启动子元件的性质详见表7-2。

表7-2 真核生物启动子元件

启动子	位置	转录因子	共有序列
Inr	+1	TBP	$PyPyA_{+1}N(T/A)PyPy$
TATA box	-35~-20	TBP	TATAAA
CAAT box	-200~-70	CBF, NF1, C/EBP	CCAAT
GC box	-200~-70	SP1	GGGCGG

注:大多数情况下,CAAT和GC box位于-200到-70的位置上。CBF为CAAT结合蛋白;C/EBP为CAAT/增强子结合蛋白。

与起始子和TATA box互作的蛋白质被称为TATA-box结合蛋白(TATA-box binding protein, TBP)。TBP是转录因子TFⅡD的一个亚基,它不仅能够识别蛋白质编码蛋白的核心启动子,同时也识别RNA启动子。

(二) 增强子

增强子是一类DNA元件,通过与转录因子(激活子)的结合,可以增强转录。它可以位

于转录起始位点的上游或下游。大多数增强子位于转录起始位点上游。在原核生物中，增强子与启动子十分接近，而真核生物的增强子可能远离启动子。

（三）沉默子

沉默子是这样一类DNA元件：它通过与转录因子（抑制子）结合，可以抑制转录。在原核生物中，沉默子也被称为操纵子（operators），可以在许多基因中找到，如lac操纵子和trp操纵子。在真核生物中，下列基因被证明含有沉默子：人β globin基因（binding of HMG-I（Y）elicits structural changes in a silencer of the human beta-globin gene.）；人CD95（Fas/APO-1）基因（silencer and enhancer regions in the human CD95（Fas/APO-1）gene with sequence similarity to the granulocyte-macrophage colony-stimulating factor promoter: binding of single strand-specific silencer factors and AP-1 and NF-AT-like enhancer factors.）；人dopamine beta-hydroxylase（DBH）基因（The cell-specific silencer region of the human dopamine beta-hydroxylase gene contains several negative regulatory elements.）以及脑源性神经营养因子基因（brain-derived neurotrophic factor expression in vivo is under the control of neuron-restrictive silencer element.）。

在少数情况下，一个DNA元件可以根据所结合的蛋白质来发挥增强子或者沉默子的作用。例如，一些基因含有一个称为E box的元件（共有序列为CACGTG），它可以与Max/Myc二聚物或者Max/Mad二聚物结合。Max/Myc二聚物可以激活转录，而Max/Mad二聚物则抑制这些基因的转录。

（四）响应元件

响应元件是一类转录因子的识别位点（表7-3）。大部分响应元件位于转录起始位点1kb范围内。

表7-3 真核生物响应元件

响应元件	转录因子	共有序列
CRE	CREB	TGACGTCA
ERE	雌激素受体（Estrogen receptor）	AGGTCANNNTGACCT
GRE	糖皮质激素受体（Glucocorticoid receptor）	AGAACANNNTGTTCT
HSE	热休克因子（Heat shock factor）	GAANNTTCNNGAA
SRE	血清应答因子（Serum response factor）	CC（A/T）$_6$GG

注：（A/T）$_6$为6个A或T。

cAMP应答元件（CRE）与CREB（CRE结合蛋白）相互作用，CREB由cAMP调控。

雌激素应答元件（ERE）和糖皮质激素应答元件（GRE）分别是雌激素受体和糖皮质激素受体的识别位点。需要注意的是，虽然激素本身不是转录因子，但是许多激素的受体却是转录因子。

血清应答元件（SRE）与血清应答因子（SRF）结合，SRF可以被许多血清中的生长因子激活。AP-1的Fos亚基由一个包含SRE的基因编码，Fos通常被认为在细胞周期过程中发挥了重要的作用。

四、真核生物转录机制

在真核生物中，DNA组装成染色质，通过限制RNA聚合酶及其附属因子与DNA的结合将基因维持在一个“失活”的状态。染色质由组蛋白构成，组蛋白形成的结构称为核小体。组蛋白可以被翻译后修饰，通过降低核小体的能力从而抑制转录因子的结合。基因的“开启”和“关闭”是一个预编程的方式，即一个最终形成细胞特异性的过程。这个编程的过程是由转录因子精心策划的，它们通常与被它们控制的基因附近的一些特殊DNA位点相结合。单个的转录因子并不能决定一个调控事件。相反，组合控制机制才是调控的关键。在组合控制中，不同组合以及细胞类型特异的蛋白质主导了基因的开关。

第二节

启动子的信息学分析

Section 2 Bioinformatics Analysis on Promoter

一、启动子识别问题

真核基因的识别问题一直是生物信息学的一个重要内容,基因启动子区的识别是完整基因结构识别中的重要一环。启动子是一段位于结构基因5' 端上游的DNA序列,能活化RNA聚合酶,使之与模板DNA准确地结合并具有转录起始的特异性。转录的起始是基因表达的关键阶段,而这一阶段的重要问题是RNA聚合酶与启动子的相互作用。启动子的结构影响了它与RNA聚合酶的亲和力,从而影响了基因表达的水平。人类启动子区的识别是生物医学研究的基本需要,是构建基因调节网络的一个核心问题。负责mRNA转录的RNA Pol Ⅱ启动子是启动子中数量最多,也是最重要的一类。

在早期的启动子预测的研究中,隐马尔科夫模型、类神经网络、数据挖掘与权重矩阵等方法被广泛应用。目前预测启动子主要从鉴定启动子的转录起始位点、核心启动子区域、转录因子结合域和启动子的CpG岛四个方面出发。但是,当用这些启动子预测工具来处理未知的、复杂的DNA序列时,识别的结果往往是比较严重的遗漏和偏高的假阳性率。

二、启动子数据资源

公共分子信息数据库包括基因图谱数据库、核酸序列数据库、蛋白质序列数据库、大分子结构数据库等。这些数据库由专门的机构建立和维护,他们负责收集、组织、管理和发布生物分子数据,并提供数据检索和分析工具,向生物学研究人员提供大量有用的信息,最大限度地满足他们的研究需要,为生物信息学研究提供服务。

目前,国际上有三个主要的核酸序列数据库: 美国国家生物技术信息中心(NIH)建立的DNA数据库, GenBank; 欧洲生物信息研究院(European bioinformatics institute, EBI)创建的核苷酸序列数据库, EMBL; 以及日本DNA数据库, DDBJ(DNA data bank of Japan),(表7-4)。这三个数据库分别在全世界范围内收集序列信息,同时,他们每天都将新发现或更新过的数据相互交换。

表7-4　国际性的核苷序列数据库

数据库	网址
GenBank	http: //www.ncbi.nlm.nih.gov/
EMBL nucleotide sequence database	http: //www.ebi.ac.uk/embl/
DNA data bank of Japan（DDBJ）	http: //www.ddbj.nig.ac.jp/

在本节中使用的启动子数据，除了可以从上文提到的三个综合性数据库中下载外，还有一些专门针对启动子数据建立的数据库（表7-5）。这些数据库通常是启动子识别研究工作者们获取启动子数据的主要来源。各个数据库的数据描述和数据的收集方法及开发工具等在其网站上均有详尽的描述，用户可以根据自己的需要来选择搜索并下载相关的数据。这些数据提供了启动子的序列信息、位置信息以及类别信息等，并且其中的部分数据库，如真核生物启动子数据库（EPD）等，还保证了所含启动子数据非冗余性。

表7-5　部分启动子/转录起始位点数据库及网址

数据库	网址
eukaryotic promoter database（EPD）	http: //www. epd. isb-sib. ch/
database of transcriptional start sites（DBTSS）	http: //dbtss. hgc. jp/
hematopoiesis promoter database（HemoPDB）	http: //bioinformatics. wistar. upenn. edu/HemoPDB
mammalian promoter database（MPromDb）	http: //bioinformatics. wistar. upenn. edu/HemoPDB
human chromosome 22 promoter data	http: //www. sanger. ac. uk/about/history/hgp/chr22. html
transcription regulatory regions database（TRRD）	http: //www. bionet. nsc. ru/trrd/
transcriptional regulatory element database（TRED）	http: //rulai. cshl. edu/cgi-bin/TRED/tred. cgi ? process=home

（一）真核基因启动子数据库EPD/EPDnew

1. 数据库概况　真核基因启动子数据库（eukaryotic promoter database，EPD）由以色列Rehovot的 Weizmann 科学研究所设计和开发。目前，由Epalingess/瑞士洛桑的ISREC管理和维护。由两个实验室协作完成的更新程序将会确保EPD中的位置参考和主数据库中序列数据的兼容性。EPD作为EMBL数据库中的一个专门的注释数据库，提供了相关真核生物启动子的信息，以帮助实验研究人员及生物信息学研究人员分析真核基因的转录信号。EPD目前的版本源于文献，以层次分类顺序组织起来，所记录的功能位点数据集指向转录起始位点。EPD中的所有信息或者直接来源于科学文献，或者从第73版本继承。因此，EPD中的启动子信息独立于EMBL序列条目描述。同样，EPD中出现的许多起始位点并不出现在相应的EMBL功能表中。EPD是目前唯一的实验证实启动子数据库，所以是各种预测软件的评论手段之一。

作为一个带有注释信息的非冗余真核生物聚合酶（POL）Ⅱ启动子数据库，EPD中的转录起始位点信息均由实验证实，如：是否为真核RNA聚合酶Ⅱ启动子、是否在高等真核生物

中有生物学活性、是否与数据库中的其他启动子有同源性等。一个条目的注释部分包括对起始位点映射数据的描述、与其他数据库的交叉引用（如EMBL、SWISS-PROT、TRANSFAC等）以及对参考文献的描述。EPD的结构及组织方式有利于动态提取有生物学意义的启动子集用于序列比较分析。截至本书编稿，该数据库已经包含了十个物种共4806条启动子序列。

EPDnew重新收集了在人类和小鼠基因组中经过实验验证的启动子。证据来自CAGE、TSS-seq等高通量实验的TSS图谱。分析时同时考虑了H2AZ、H3K4me3、POLⅡ及DNA甲基化的ChIP-seq实验结果。数据库最终包含9716个人类启动子和9773个小鼠启动子。

2. 启动子数据的检索　用户可以直接从EPD网站的FTP站点批量下载Fasta格式的启动子数据。也可以在Download EPD db中，依照所需启动子的位置及大小获取数据。如图7-5中选择的是TSS上游-499bp至下游100bp，长度为600bp的启动子序列。

3. 数据的格式　EPD中的数据采用本质上相同的两种ASCII格式（epd.dat，epd_bulk.dat）存储。EPD文件包含一个标题行（图7-6），随后记录了一系列的启动子数据。为了使整个数据库与现有信号搜索分析软件使用的标准FTP文件格式一致，标题行和启动子数据的部分子项均有固定的格式。

每条启动子数据的存储格式与EMBL及SWISS-PROT序列数据的存储格式相似。每行

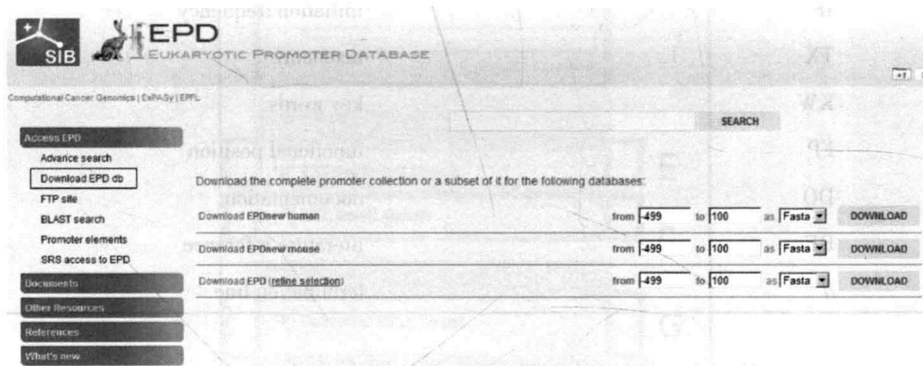


图7-5　特定位置启动子数据的下载

```
TI   EPD83    Eukaryotic Promoter Database / Release 83                EP
```

图7-6　EPD数据标题行示例

开始是一个行标，定义本行所表述信息的类别。行标的意义见表7-6。

表7-6　EPD数据行Code对应意义

Code	解释
ID	identification
AC	accession number(s)
DT	Date
DE	description

续表

Code	解释
OS	organism species
HG	homology group
AP	alternative promoter
NP	neighbouring promoter
DR	database cross-references
RN	reference number
RX	Reference cross-references
RA	reference authors
RT	reference title
RL	reference location
ME	methods
SE	sequence
FL	full length
IF	initiation frequency
TX	taxonomy
KW	key words
FP	functional position
DO	documentation
RF	literature reference
//	termination line

其中每个条目具体的解释可以在数据库网站提供的用户手册中查找到。

【例7-1】试从EPD数据库中查找到10条人类启动子数据，启动子的大小为转录起始位点上游 -1300bp至TSS下游 +49bp。利用该网站提供的比对工具进行BLAST比对。

解答：①登录EDP数据库：http: //epd.vital-it.ch/；②从数据库页面左侧Access EPD功能中选择Download EPD db；③在Download EPDnew human中将所需启动子范围定在-1300bp到49bp，点击download；④选择10条满足条件的fasta格式序列在blast工具中比对即可得到比对结果。

（二）转录起始位点数据库DBTSS

1. 数据库概况　DBTSS（database of transcriptional start sites）是东京大学医学科学院人类基因组中心（human genome center，institute of medical science，The University of Tokyo）开发的一个关于启动子及转录调控的研究数据库。其中包含了精确的真核生物mRNA转录起始位点信息。在最近版本中，数据库增加了新的TSS数据，使数据库覆盖了大部分成人及胚胎组织。目前，DBTSS包含收集自20个组织及7个细胞系的49，100万条TSS标签序列。数据库

还整合了最近产生的RNA-seq数据及组蛋白修饰的ChIP-seq数据。用户不仅可以得到精确的TSS位置信息，还能得到它们的表达水平，这有助于进一步推断启动子上游区域并理解基因的转录调控。

2. 数据库检索 DBTSS的主要有个四个部分构成，分别是：数据库搜索模块、TSS-Seq/SNP信息搜索模块、分析工具模块及下载模块。在其网站主页的左边栏可以清晰地看到各个模块的构成及功能（图7-7）。

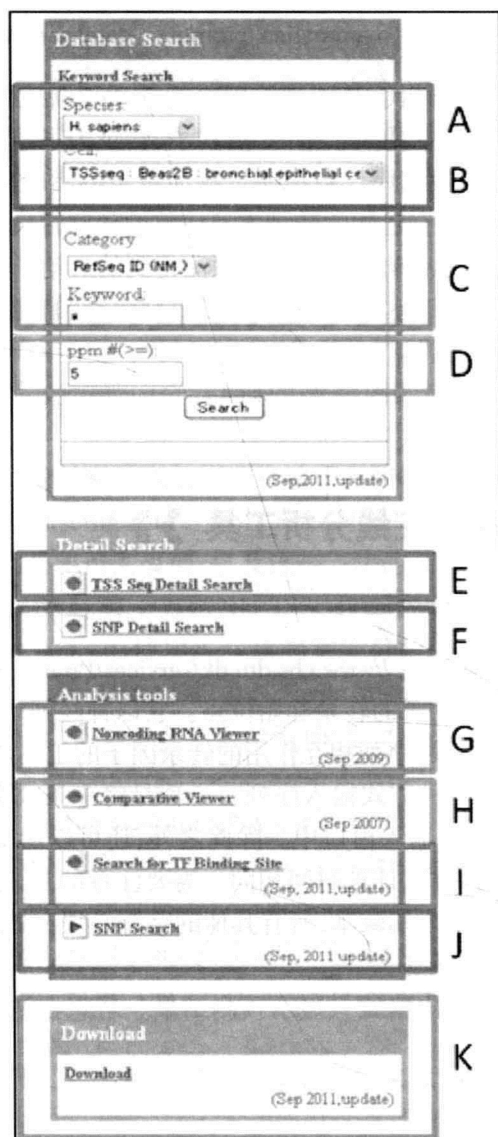


图7-7 DBTSS数据库的构成

A. 选择所要搜索的物种；B. 搜索TSS-Seq数据时，选择“TSS Seq”；搜索cDNA数据时，选择“Sanger”；C. 输入查询条件；D. 搜索TSS-Seq数据时，输入最小ppm；搜索cDNA数据时，输入克隆/标签号；E、F.TSS-Seq/SNP信息搜索模块；G. 非编码RNA浏览器；H. 比较浏览器；I. 转录因子结合位点搜索工具；J. SNP搜索工具；K. 数据库下载FTP

用户可以根据不同的需求，使用不同的搜索及分析工具。

【例7-2】试在DBTSS数据库查找CTCF（CCCTC）在人类MCF7细胞系中可能的转录因子结合位点。

解答：①登录到DBTSS数据库：http://dbtss.hgc.jp/；②在数据库左侧工具栏中Analysis Tools中点击Search for TF Binding Site工具；③选择目标物种和细胞系：人类MCF7，设置转录因子位置，点击create selection table按钮；④在模式输入框中，填入所查找的CTCF的结合motif：CCCTC，点击搜索，即可得到可能的转录因子结合位点信息。

（三）哺乳动物启动子数据库（mammalian promoter database，MPD）

数据库概况：在后基因组时代，基因调控网络的性质逐渐成为基因组研究中重要的一部分。MPD就是在这个时候建立的关于基因、启动子、转录因子结合位点以及其他顺式调控元件的一个高质量且全面的数据库。数据库全称为冷泉港实验室哺乳动物启动子数据库（CSHLmpd）。数据库使用了所有已知的转录本及完整的预测转录本，构建了人类、大鼠和小鼠基因组的基因集合。其中的启动子信息包含预测得到的启动子。数据库中的启动子全部映射到基因组中，和相关基因相连。数据库还对垂直同源基因组的启动子进行了比较分析，以检测启动子区域序列的保守性。

使用CSHLmpd有助于基因调控网络的研究，它向如DNA microarray等实验提供研究指导。

三、真核生物启动子在线分析工具

（一）Promoter 2.0 Prediction Server

Promoter2.0预测服务器（http://www.cbs.dtu.dk/services/Promoter/）的主要功能是在DNA序列中预测脊椎动物POLII启动子的转录起始位点。它以神经网络和遗传算法为基础，已经发展成为一个模拟在启动子区域序列相互作用的转录因子的工具。

1.输入序列的处理　有两种方式输入序列。一种是将一条或多条FASTA格式序列直接粘贴到服务器主页上部的序列输入窗口中。除此之外，还可以从本地硬盘中选择待处理的FASTA文件，直接上传。两种方式计算时间相同。需要注意的是，序列文件中所用字符必须为A，C，G，T或X。其中X代表未知碱基，所有其他的字符必须在处理前转换成X。单次输入限制为最多50条序列或150万碱基。

2.选择输出格式　默认输出格式只显示预测结果。若想在结果中包含输入序列，则需要点击“Full output”按钮。

（二）PromoSer

人们对转录调控机制研究的关注点一般在基因启动子区域附近。人们需要获取这些区域来寻找大量相关基因。用计算的方法来预测整个基因组的启动子及基因在很大程度上依赖于训练模型时使用的预先确定的数据集。这就需要收集大量高精确度的启动子序列。

PromoSer（http://cagt.bu.edu/page/Promoser_submit）[6]是一个基于网络的服务，旨在提取大量哺乳动物基因组启动子序列。为了识别一个基因的转录起始位点（TSS），创立者将所

有可用的mRNA及EST序列数据映射到基因组中，通过跟踪重叠比对，获得这些序列最大的延伸可能，最终确定TSS。

PromoSer易于使用。只要提供一个GenBank登录ID列表，以确定感兴趣的基因，并输入所需TSS的侧翼范围。PromoSer处理输入并返回一个含有所需区域的多重FASTA格式文件。

(三) neural network promoter prediction, NNPP

NNPP(http: //www.fruitfly.org/seq_tools/promoter.html)是一个在DNA序列中发现真核及原核生物启动子的方法。NNPP程序以一个时间延迟神经网络为基础。时间延迟神经网络主要包含两个功能层：一个用来识别TATA-box，另一个识别“起始子”，即一段包含转录起始位点的区域。两个功能层合并成一个输出单元，输出的得分在0~1之间。

【例7-3】在GenBank中查找一条真核生物DNA序列，利用真核生物启动子在线分析软件预测分析启动子区域，并分析启动子特征。

解答：用前文介绍的在线分析软件做实验练习。

四、启动子识别的信息学研究方法

由于启动子在基因转录过程中发挥着至关重要的调控作用，使得对启动子的识别研究成为科学研究者关注的焦点。

从研究的方法上看，目前已经发表的用于启动子预测程序的识别技术有基于神经网络、线性和二次判别(quadratic discriminate)分析、相关向量机(relevance vector machine)、启动子区域的统计性质、改进的马尔可夫模型，以及这些方法的结合。

比如常用的计算机预测启动子方法中，TSSG和TSSW使用了位点比重阵列(position weight matrix, PWM), Core Promoter使用了二次判别式分析(quadratic discriminant analysis, QDA)；基于隐马尔可夫模型(HMM)的方法Audic和Mcpromoter中，Audic是HMM结合贝叶斯定律，Mcpromoter使用了HMM结合高斯分布曲线；而DranonPF、DragonGSF、NNPP2.2、Promoter2.0用人工神经网络(artificial neural network, ANN)作为方法设计的一部分，Promoter2.0使用遗传算法(genetic algorithm, GA)，而NNPP又结合了时间延迟神经网络(time-delay neural network, TDNNS)和位点修剪。除了计算及统计方法上的不同之外，一些方法还对序列本身的性质加以利用，比如PromFind和PromoterInspector应用了六聚体和低聚复合物的性质，DragonPF、DragonGSF、Eponine和FirstEF考虑了G + C的含量；Eponine、NNPP2.2和Promoter2.0引入了“TATA盒”模块；CpGProD、DragonGSF和FirstEF从不同角度对CpG岛提供的信息加以利用。

基于对以上方法的了解，我们可以将用计算的方法来预测识别启动子的方法大致分成三类：一类是基于统计或内容的方法，这类方法通过计算低聚核苷酸的重复频率或比较转录因子结合位点出现的频率来对启动子进行分析。第二类是基于神经网络的方法，这类方法使用了ANN这种信息处理系统对启动子进行识别。大部分的方法属于第三类，即对前两类技术的结合利用。以Dragon Promoter Finder(DragonPF)为例，DragonPF是一个预测脊椎动物启动子的综合的启动子预测模型。它结合了对一段未知序列依次进行五聚体筛选、PWM位点分析、信号处理以及人工神经网络的方法。目前启动子识别问题的发展趋势是将启动

子和编码外显子以及内含子等非启动子序列同时加以考虑,并用合成的方法来分析。

从研究方法所着眼解决的问题上对启动子识别问题进行分类: 随着识别方法研究的不断发展,对启动子序列的研究渐渐主要分成解决识别一段序列是否是启动子以及在识别的同时确定一段基因的转录起始位点两大类问题。

目前常用的RNA POL Ⅱ 启动子识别方法见表7-7,常用的转录起始位点识别方法见表7-8,另外,还有一些一般的基因识别方法可以用来进行RNA POLⅡ 以及其他特征(MARs、CpG岛)的检测,见表7-9。

表7-7　常用的RNA POL Ⅱ启动子识别方法

方法名称	相关网站及信息
Audic/Claverie	向audic@newton.cnrs-mrs.fr发送请求
CorePromoter	http: //rulai.cshl.edu/tools/genefinder/CPROMOTER/index.htm
FunSiteP	http: //compel.bionet.nsc.ru/FunSite/fsp.html
ModelGenerator/ModelInspector	http: //www.gsf.de/ieg/groups/biodv/modyproject.html
PPNN	http: //www.fruitfly.org/seq_tools/promoter.html
PromFD 1.0	FTP
PromFind 2.0	http: //www.rabbithutch.com/
Promoter 2.0	http: //www.cbs.dtu.dk/services/Promoter/
Promoter Scan	http: //thr.cit.nih.gov/molbio/proscan/
TSSG/TSSW	http://www.softberry.com/berry.phtml ? topic=index&group=programs&subgroup=promoter

表7-8　常用转录起始位点识别方法

方法名称	相关网站及信息
MatInd/MatInspector/FastM	http: //www.gsf.de/ieg/groups/biodv/modyproject.html
MATRIX SEARCH 1.0	向chenq@boulder.colorado.edu发送请求
PatSearch 1.1	http: //www.800xe.de/webwatch/Patsearch-Das-private-Experiment.html
Signal Scan	http: //bimas.dcrt.nih.gov/molbio/signal/
TESS	http: //www.cbil.upenn.edu/tess/
TFSEARCH	http: //www.cbrc.jp/research/db/TFSEARCH.html

表7-9　用于分析RNA POLⅡ的基因识别方法

方法名称	相关网站及信息
GENSCAN	http: //genes.mit.edu/GENSCAN.html
GRAIL	http: //compbio.ornl.gov/Grail-1.3/
MAR-Finder	http: //www.futuresoft.org/MAR-Wiz/

续表

方法名称	相关网站及信息
WebGene	http: //www.itb.cnr.it/sun/webgene/
GENSCAN	http: //genes.mit.edu/GENSCAN.html
GRAIL	http: //compbio.ornl.gov/Grail-1.3/

从研究者使用的数据上分类,一部分人选择广义概括的启动子序列进行研究,如真核生物启动子或脊椎动物启动子序列;而另一部分人则选择使用更特殊化的启动子,如大肠埃希菌启动子。

但是,由于:①核心启动子(core promoter)并不是一个单一的类型;②启动子序列之外还有许多额外的调整元素;③转录过程可能被规则的蛋白质(regulatory proteins)活化或抑制;④转录催化剂和抑制剂有特定的作用并且与细胞类型和在细胞周期中的点均有关系等原因,使得对启动子的识别仍然是一项很艰难的工作。现已发表的对不同类型生物或物种启动子识别的方法仍然很难达到一个好的准确度。由于分析问题的复杂程度以及解决问题的方法的不同,现有方法对整个人类基因组启动子识别的准确率仅为50.00%左右,而对特定种类的启动子数据的识别准确率在85.00%~86.00%之间,并且仍然存在相当高的假阳性率。

第三节

转录因子结合位点的信息学研究

Section 3 Bioinformatics Analysis on TF Binding Site

一、转录因子及其转录调控机制

在分子生物学和遗传学中，一个转录因子（有时被称为序列特异的DNA结合因子）是一个能与特异DNA序列结合的蛋白质。转录就是调控遗传信息从DNA传递到mRNA的过程。转录因子可以单独或与其他蛋白质形成复合体，提高或阻断特异基因对RNA聚合酶的招募。

转录因子的一个特点是它包含一个或多个DNA结合域（DNA-binding domain, DBDs），通过这些结合域与基因附近的DNA序列结合，从而完成调控。其他蛋白质，如共激活因子（coactivators），染色质重构因子（chromatin remodelers），组蛋白乙酰化酶（histone acetylases），去乙酰化酶（deacetylases），激酶（kinases）和甲基化酶（methylases），虽然在基因调控中同样起着重要作用，但是由于缺少DNA结合域，因而并没有被归类为转录因子。

（一）转录因子在不同生物中的保守性

转录因子存在于所有生物体中，对基因表达调控来说是必不可少的。在一个生物体内转录因子的数量随着基因组大小的增加而增长，较大的基因组每个基因倾向于有更多的转录因子。在人类基因组中大约有2600个含有DNA结合结构域的蛋白质，其中大多数被假设具有转录因子功能。因此，基因组中大约10%的基因编码转录因子，这使得这个家族成为最大的人类蛋白质家族。此外，基因的两侧往往存在不同的转录因子结合位点，这些基因的高效表达需要几个不同的转录因子的协同作用。

（二）转录因子调控机制

转录因子可以与受其调控基因临近DNA上的增强子或启动子区域结合。根据转录因子的不同，相邻基因的表达可能被上调或下调。转录因子有多种调控基因表达的机制，包括：

1. 稳定或组织RNA聚合酶与DNA结合。

2. 催化组蛋白的乙酰化或脱乙酰化　转录因子可以直接或招募其他带有这一催化活性的蛋白质来完成这一作用。许多转录因子使用两种对立机制的其中一种来调控转录：①组蛋白乙酰转移酶作用——组蛋白乙酰化，从而削弱了DNA与组蛋白的结合，使得DNA更容易转录，起到转录上调的效果；②组蛋白去乙酰化酶（histone deacetylase, HDAC）作用——组蛋白去乙酰化，从而加强了DNA与组蛋白的结合，使得更少的DNA暴露，达到下调转录的目的。

3. 为转录因子DNA复合物招募共激活子或共抑制子蛋白质　在生物学中，重要的进程大多具有多层调控的性质。转录因子也具有这样的性质：转录因子不仅能通过控制转录率来调节细胞中基因产物（RNA和蛋白质）的数量，而且转录因子自身也受到调控作用（通常被其他转录因子调控）。下面将对转录因子调节方式及活性做简短的概述：

（1）合成：转录因子（像所有蛋白质一样）从一个染色体上的基因转录成RNA，然后被翻译成蛋白质。调控这些步骤中的任何一步都会影响转录因子的产生（以及活性）。这里存在一个有趣的现象，即转录因子可以被自己调控。例如，在一个负反馈环中，转录因子作为自己的抑制子：如果转录因子蛋白质与自身基因的DNA结合，它将会抑制自身的产生。这是一类在细胞中转录因子能够维持较低水平的机制。

（2）核定位：在真核生物中，转录因子（像大多数蛋白质一样）在细胞核中转录，但是在细胞质中翻译。许多在细胞核中具有活性的蛋白质含有核定位信号，能直接定位细胞核。但是，对许多转录因子来说，这是在它们调控过程中的关键。几类重要的转录因子，如一些核受体转录因子在细胞质中必须先在绑定一个配体，才能重新定位细胞核。

（3）激活：转录因子可以通过它们的信号感应区域激活（或失活），机制包括：

1）配体的结合：配体结合不仅能够影响转录因子在细胞内的位置，也可以影响转录因子是否处于激活的状态，从而能够与DNA其他辅助因子结合（例如，核受体）。

2）磷酸化：许多转录因子，如STAT蛋白只有磷酸化后才能与DNA结合、与其他转录因子或共调控蛋白相互作用（如，同源或异二聚体）。

4. 易接近DNA绑定位点　在真核生物中，DNA在核小体的帮助下组织成压缩的状态，其中约147个DNA碱基对在组蛋白八聚体周围缠绕两圈。核小体内部的DNA无法与转录因子接近。一些转录因子，被称为先锋因子，仍然能够与核小体DNA的DNA绑定位点结合。对大多数其他转录因子来说，核小体必须被如染色质重塑子等分子驱动零件激活转移。另外，核小体可以被热波动部分解开，使得转录因子结合位点暂时性暴露出来。在许多情况下，一个转录因子与DNA绑定位点的结合需要与其他转录因子、组蛋白或非组蛋白染色质蛋白质进行竞争。转录因子与其他蛋白的组合在调控相同的基因上可以发挥相反的作用（激活与阻遏）。

5. 其他辅助因子/转录因子的可用性　大多数转录因子不单独工作。通常情况下，为了完成基因的转录，一系列的转录因子必须与DNA调控序列绑定。这种转录因子的集合，反过来，招募中介蛋白质，如cofactor，以高效招募前起始复合物和RNA聚合酶。因此，对于一个单一的转录因子起始转录，所有这些其他蛋白质也必须在场，并且，这个转录因子必须处于一旦需要就能够结合到这些蛋白质的状态中。

（三）转录因子的功能

转录因子是这样一组蛋白质，它们阅读和诠释DNA中的遗传“蓝图”。它们与DNA结合，并帮助启动一个负责基因转录增加或减少的程序。因此，它们对许多重要的细胞过程来说是至关重要的。下面是一些转录因子参与的重要功能和生物学角色：

1. 基础转录调控　在真核生物中，又称为一般转录因子（general transcription factors，GTFs）的一类重要的转录因子，它们是发生转录的必要条件。这些GTFs中，很多实际上都不绑定DNA，而仅仅作为大转录前初始复合物的一部分，与RNA聚合酶直接互作。最常见的

GTFs是TFⅡA、TFⅡB、TFⅡD、TFⅡE、TFⅡF和TFⅡH。起始前复合物与其调控基因上游的启动子区域DNA相结合。

2. 转录的差异性增强 有部分转录因子可以绑定邻近调控基因的DNA增强子区域，从而差异调控各种基因的表达。对保证基因在适当的时间、适当的细胞适量的表达以适应有机体不断变化的需求，这些转录因子起到了至关重要的作用。

3. 发育 在多细胞生物体内，许多转录因子参与到发育的过程中。这些转录因子调控相应基因的表达与否，决定细胞的分化、细胞形态或活性的变化。以Hox转录因子家族为例，对于有机物（如从人类到果蝇的多样化）正确的体型形成十分重要。另一个例子是由性别决定区域Y（SRY）基因编码的转录因子，在决定人类性别的过程中发挥了重要作用。

4. 细胞信号的响应 通过释放一种可以产生与受体细胞进行信号传导的分子，使细胞之间可以互相沟通。如果这个信号需要上调或下调受体细胞内的基因表达，那么转录因子将会在信号级联的下游出现。雌激素信号是一个与雌激素受体转录因子有关的相当短的信号级联的例子：雌激素由组织（如卵巢和胎盘等）分泌出来，穿过受体细胞的细胞膜，在细胞质中与雌激素受体结合。接着，雌激素受体进入细胞核，绑定到DNA绑定位点，改变相关基因的转录调控。

5. 环境应答 转录因子不仅能在生物刺激有关的信号级联下游发挥作用，同时，它们也能参与环境刺激的信号级联下游。例如热休克因子（heat shock factor，HSF），它可以上调在高温下耐受的必需基因；缺氧诱导因子（hypoxia-inducible factor，HIF），它能够上调在低氧环境中生存的必需基因；以及胆固醇调节元件结合蛋白（sterol regulatory element binding protein，SREBP），它能有助于维持细胞的正常血脂水平。

6. 细胞周期控制 许多转录因子，特别是一些原癌基因或肿瘤抑制基因，有助于调节细胞周期和决定当一个细胞长到多大时分裂成两个子细胞。以致癌基因Myc为例，它在细胞增长和凋亡过程中有重要的作用。

7. 发病机制 转录因子也可用于改变宿主细胞的基因表达，促进发病机制。由黄单胞菌分泌的类似转录激活子的作用因子（TAL effector）就是这方面一个被广泛研究的例子。当这些蛋白被注射到植物中，它们能够进入植物细胞的细胞核中，与植物启动子序列结合，激活帮助细菌感染的植物基因的转录。

（四）转录因子的结构

转录因子具有模块状结构（图7-8），包含如下结构域：

DNA结合结构域（DNA-binding domain，DBD），与被调控基因相邻的特定DNA序列（增强子或启动子）相结合。能与转录因子结合的DNA序列通常被称为响应元件。

反式激活结构域（trans-activating domain，TAD），其中包含其他蛋白质（如转录共调控子）的结合位点。这些结合位点通常被认为具有激活功能（AFs）。

一个可选的信号感应结构域（signal sensing domain，SSD）（例如，一个配体结合域），它可以感应外部信号，并且将这些信号传导到其余的转录复合物，导致基因表达的上调或下调。此外，DBD和信号感应结构域可以存在于不同的蛋白质中，在转录复合物中相互作用，完成对基因表达的调控。



图7-8　转录因子的模块状结构

一个转录因子氨基酸序列(N端在左C端在右)示意图。包含一个DNA结合域(DBD);信号感知域(SSD)和一个反式激活结构域(TAD)。在不同转录因子中这些结构域的顺序和数量是不同的。另外,反式激活结构域和信号感知结构域的功能通常包含在相同的结构域中

二、转录因子结合位点的高通量试验技术

(一) 染色质免疫沉淀芯片(Chip-chip)

该技术能够快速在目标基因组的染色体中确定特异DNA结合蛋白的准确结合位点,ChIP芯片也可以在一个基因组的任何感兴趣的区域内寻找染色体的结构改变。

1. ChIP-chip的用途

(1)在基因组范围内确定基因转录因子的DNA结合位点和其他DNA结合蛋白或蛋白复合体的DNA结合位点。

(2)染色体活性状态的定量分析。

(3)组蛋白修饰的功能研究。通过用酰基化或甲基化的组蛋白的特异抗体和没有进行修饰的组蛋白的特异抗体,可以确定与组蛋白修饰有关的结合模式的变化。

(4)聚合酶活性的定量分析。

(5)精炼生物信息方法,用功能数据来确定启动子的位置。

2. 具体实验原理和实验步骤如图7-9。

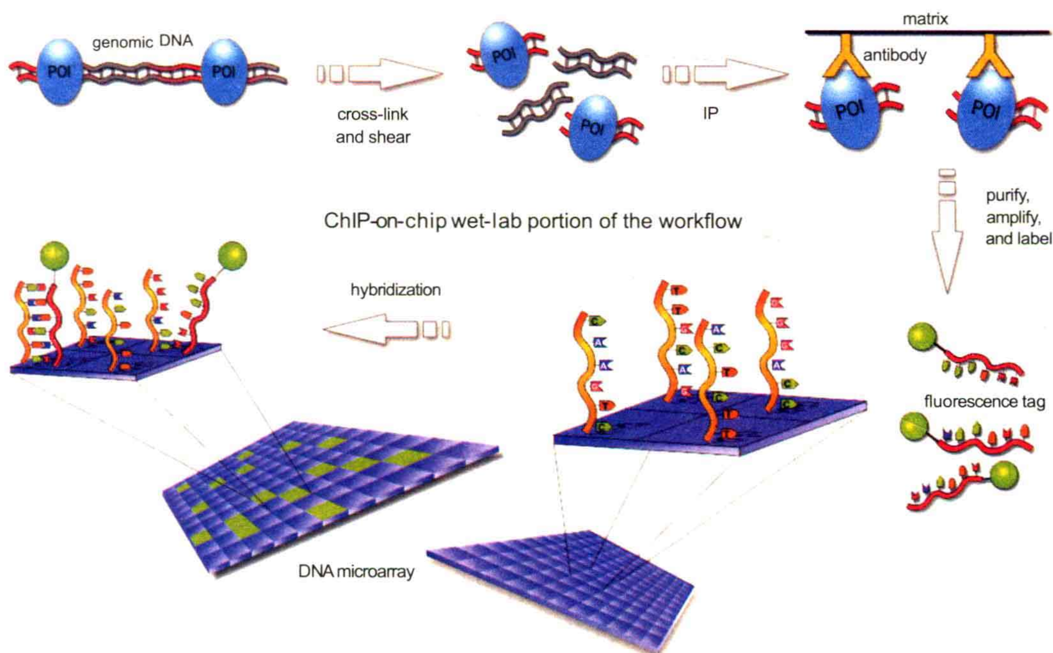


图7-9　染色质免疫沉淀芯片流程图

3. GeneChip-TilingArray技术简介

Affymetrix公司于2006年1月24日宣布推出GeneChip(R)人类和鼠源嵌合芯片(TilingArray)系列产品。该系列芯片研究范围大大超出已知编码蛋白序列,可以对整个人类和小鼠基因组进行系统的研究。研究人员可以利用这一芯片对转录因子和其他蛋白结合结构域进行研究。最近,更有研究人员利用Affymetrix的嵌合芯片在过去认为是垃圾DNA的区域中间找到了许多以前从未发现过的转录活性区域。嵌合芯片(TilingArray)是迄今为止分辨率最高的基因芯片类型,其探针设计几乎涵盖了目标DNA的全部序列。迄今为止,Affymetrix公司已经开发出了人、小鼠、酵母、线虫、拟南芥等模式生物的全基因组Tiling芯片,为全基因组规模上研究目的蛋白与核酸的相互作用提供了强有力的分析工具。GeneChip-TilingArray除了全基因组芯片外,还包括了专门应用于ChIP—chip技术中的人启动子和小鼠启动子两款芯片,探针设计覆盖了转录起始位点附近10kb的范围,可针对肿瘤相关的1300个基因,覆盖范围更是增加到了12.5kb。

1882年,德国细胞学家弗莱明首次公开发表了细胞有丝分裂现象的观察结果,他的工作也被看做是科学史上最重要的发现之一。除了对有丝分裂进行描绘以及命名之外,弗莱明还对这一过程中看似起关键作用的物质——染色质作了标记。

目前,它是生物学两个最热门领域——基因组学和蛋白组学研究关注的焦点。但是,不同之处在于:弗莱明采用的是光学显微镜和装有少量苯胺染色的玻璃瓶对其进行研究,而最新的基因组阶段的染色质研究采用的是尖端的技术——染色质免疫沉淀作用测定法(ChIP)。

基因组学和蛋白组学都将把染色质作为研究对象,但两个领域采用的方法各异。在基因组学研究中,研究人员通常从一个蛋白质开始研究,采用ChIP去找出与基因组关联的蛋白质。而蛋白组学研究采用的是反向方法,先用一个特殊的DNA序列作为寻找蛋白质的诱饵;然后用ChIP去证实:那些蛋白质就是在体内与DNA序列相关联的蛋白质。

(二)染色质免疫沉淀-测序(ChIP-Seq)

ChIP-Seq,即染色质免疫共沉淀-测序技术,是通过对染色质免疫共沉淀(ChIP)获得的DNA片段后进行大规模测序,从而得到目标蛋白结合的DNA序列信息,并定位到全基因组上。染色质免疫共沉淀技术(chromatin immunoprecipitation , ChIP)也称结合位点分析法,是研究体内蛋白质与 DNA 相互作用的有力工具,通常用于转录因子结合位点或组蛋白特异性修饰位点的研究。将 ChIP 与第二代测序技术相结合的 ChIP-Seq 技术,能够高效地在全基因组范围内检测与组蛋白、转录因子等互作的 DNA 区段。

ChIP-Seq 的原理是:首先通过染色质免疫共沉淀技术(ChIP)特异性地富集目的蛋白结合的 DNA 片段,并对其进行纯化与文库构建;然后对富集得到的 DNA 片段进行高通量测序。研究人员通过将获得的数百万条序列标签精确定位到基因组上,从而获得全基因组范围内与组蛋白、转录因子等互作的 DNA 区段信息。

1. ChIP-Seq实验流程(以Solexa为例)(图7-10):

(1)测序:对客户提供的 ChIP 样品(如果有阴阳参启动子区域或 DNA 序列的)进行定量检测,检测合格后进行测序文库构建、DNA 成簇(Cluster generation)扩增、高通量测序。

(2)基本数据分析数据产出统计:对测序结果进行图像识别(Base calling),去除污染及

接头序列；统计结果包括：测定的序列（Reads）长度、Reads 数量、数据产量。

（3）高级数据分析（图7-11）：标准高级数据分析内容包括：ChIP-Seq 序列与参考序列比对；Peak calling：统计样品 Peak 信息（峰检测及计数、平均峰长度、峰长中位数）；统计样品 Uniquely mapped reads 在基因上、基因间区的分布情况及覆盖深度；给出每个样品 Peak 关联基因列表及 GO 功能注释；在多个样品间，对与 Peak 关联基因做差异分析。

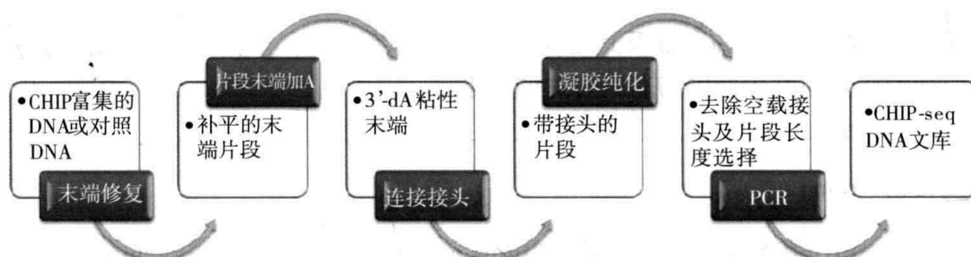


图7-10 ChIP-seq实验流程示意图

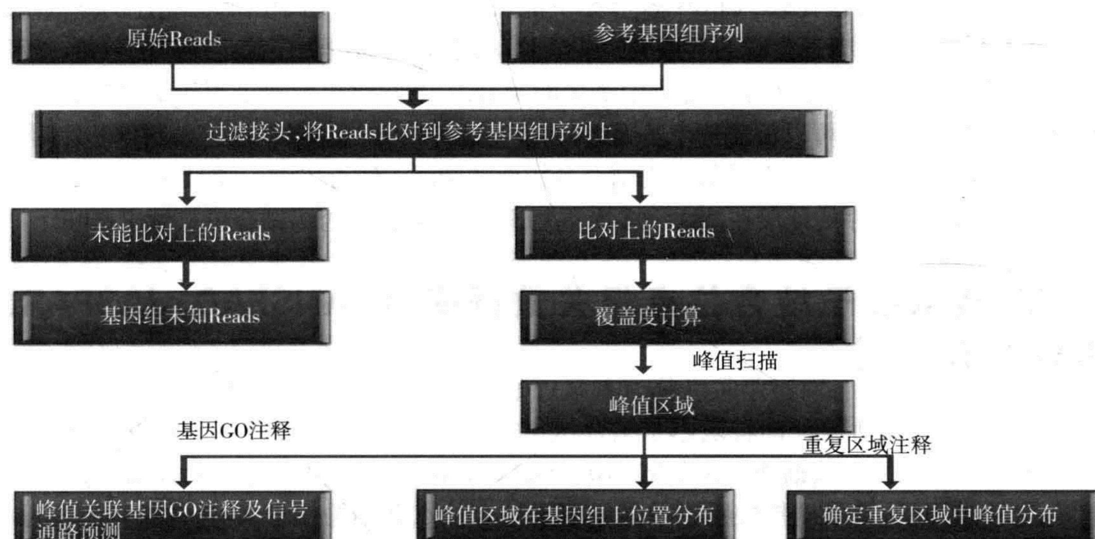


图7-11 CHIP-seq生物信息分析流程示意图

2. ChIP-Seq技术优势

（1）高通量：一个lane产生的数据几乎可以涵盖转录因子在基因组上的全部结合区域。

（2）低成本：单个read的测序和分析费用仅为传统测序法的1/100；只有全基因组ChIP-chip的1/30到1/10。

（3）灵活度高：任何物种任何序列都可进行实验，无需已知的基因组序列信息。

（4）高可信度：比ChIP-Chip更低背景水平和高信噪比确保高可信度的实验结果。

（5）信噪比高：背景比芯片结果（ChIP-chip）低，每个ChIP样本可获取数百万个有效序列标签，利用数百万次计数将真实事件与假信号区分开。

（6）检测范围广：在整个基因组内定位体内结合位点，包括芯片无法检测的重复序列

区域。

（7）定位精确度高：在50 bp以内定位结合位点。

3. ChIP-Seq应用领域

由于 ChIP-Seq 的数据是 DNA 测序的结果，为研究者提供了进一步深度挖掘生物信息的资源，研究者可以在以下几方面展开研究：

（1）判断 DNA 链的某一特定位置会出现何种组蛋白修饰。

（2）检测 RNA polymerase Ⅱ 及其他反式因子在基因组上结合位点的精确定位。

（3）研究组蛋白共价修饰与基因表达的关系。

（4）CTCF 转录因子研究。

（三）ChIP-tiling

ChIP-tiling将ChIP技术与叠片式芯片（tiling array）相结合，它与ChIP-chip不同之处就在于所采用的芯片不同。叠片式芯片对基因组的覆盖率更高。ChIP-seq将ChIP技术与高通量的测序技术相结合，ChIP实验中得到的与转录因子结合和未结合的片段可以直接被测序。与基于生物芯片的方法比较，ChIP-seq有如下优点：①它可以应用于所有已测序的基因组，而不要求有设计好的普通生物芯片或叠片式芯片；②它直接通过测序确定DNA的数量，从而避免了DNA序列与生物芯片杂交过程中产生的噪音； ③它测出的转录因子结合位点是真正无偏的（测序的方法可以覆盖整个基因组，而生物芯片的方法却限于被选入制作芯片的序列集合）；④它的灵敏度更高，能够获得结合量较低转录因子结合位点。同时，ChIP-seq亦有不足之处，如果被测序的片段在基因组多次重复，则无法对其是否为结合位点做出推断。

三、转录因子结合位点相关数据库：TRANSFAC、JASPAR、SELEX_DB

随着生物实验所验证的转录因子结合位点的不断积累，目前出现了专门收集TFBS相关信息且各具特色的数据库，详见表7-10。TRANSFAC是真核生物转录调控信息的数据库，包含转录因子，转录调控关系以及转录因子结合位点等相关信息，涵盖的物种有酵母、拟南芥、线虫、果蝇、大鼠、小鼠、人等。它通过文献挖掘来收集数据，并有严格的质量控制。TRANSFAC中收录的TFBS都是经过实验验证的，并且在每一个结合位点的条目中都标注了相应的实验技术，实验条件并对该TFBS的可信度进行了评价。TRANSFAC中不仅有TFBS的标注，还提供了相应转录因子与靶基因的信息，如物种、蛋白质一级序列、蛋白质功能域等。TRANSFAC 11.3中，共收集了10 018个转录因子，以及20 431个转录因子结合位点，为TFBS预测算法提供了高质量的训练集和验证集。

表7-10 转录因子结合位点数据库

数据库	网址
TRANSFAC	http: //www.gene-regulation.com
JASPAR	http: //jaspar.cgb.ki.se
SELEX_DB	http: //wwwmgs.bionet.nsc.ru/mgs/systems/selex/

续表

数据库	网址
HTPSELEX	http: //www.isrec.isb-sib.ch/htpselex/
PlantTFDB	http: //planttfdb.cbi.pku.edu.cn
AGRIS	http: //arabidopsis.med.ohio-state.edu
SCPD	http: //rulai.cshl.edu/SCPD
TRED	http: //rulai.cshl.edu/TRED
ITFP	http: //itfp.biosino.org/itfp

JASPAR收录了多细胞真核生物转录因子结合位点的信息,并以矩阵的形式保存,这些矩阵是由实验验证的结合位点统计得来的。JASPAR包括3个子库, JASPAR CORE、JASPAR FAM、JASPAR PHYLOFACTS。目前, JASPAR CORE中包含123个频数矩阵,矩阵中的元素表示某个位置上出现某个碱基的频数, JASPAR FAM中将转录因子按其DNA结合域的结构特性分成若干家族,并提供了11个"家族共有"的TFBS的位置权重矩阵,为从结构角度进行TFBS研究提供了方便, JASPAR PHYLOFACTS中包含174个从在进化上保守的基因上游元件中提取的频数矩阵。值得一提的是,与商业数据库TRANSFAC不同, JASPAR是完全开放的资源, JASPAR与TRANSFAC的另一个主要区别是, JASPAR中含有的TFBS信息是非冗余的,即一个转录因子对应至多一个TFBS条目。

SELEX_DB和HTPSELEX中收集了经SELEX实验验证的TFBS信息。它们不同于综合型的数据库,除了实验验证的结合位点信息,还尽可能详尽的提供了实验中间产物。此类数据库包含的TFBS相对较少,但针对每一个TFBS提供了更为丰富的实验信息,这为致力于建立更精准TFBS模型的研究者提供了宝贵的数据。

另外,还有一些收集特定物种转录因子以及TFBS信息的数据库: PlantTFDB中包含22种植物中的26 402个转录因子的信息, AGRIS中包含了模式生物拟南芥的转录因子及其结合位点的信息, SCPD是收集酵母启动子区域序列的数据库,里面包含转录起始位点以及转录因子结合位点的注释, TRED是收集哺乳动物转录调控元件的数据库,对人、小鼠、大鼠等物种的启动子区域有相对完整的注释, ITFP中收集了哺乳动物的转录因子与靶基因之间的调控关系信息。

四、转录因子结合位点模型的建立及分析

最基本的TFBS模型是一致性序列,即对结合位点中每个位置选择一个最可能出现的核苷酸组成一个序列来表达TFBS。比如某个转录因子有5个结合位点TACGAT、TATAAT、GATACT、TATAGA、TATGTT,那么它的一致性序列就是TATAAT。这样的表达方式既牺牲了特异性,也丢失了敏感性。描述TFBS的一个常用模型是位置权重矩阵模型(position weight matrix,PWM)。如果TFBS长度为L,PWM就是一个4行L列的矩阵,这个矩阵中每行对应着一种核苷酸,每列对应着TFBS中的一个位置,第i行第j列的元素是TFBS中第j位上出现核苷酸i的概率。一个长度为L的序列与该转录因子结合的概率即

为各个位置上核苷酸对应概率的乘积。某段序列与转录因子结合的概率越大，就说明它与转录因子相结合的结合能力越强。这个模型有两个局限：一是该模型中TFBS的长度是固定不变的；二是该模型假定TFBS不同位置间相互独立，每个核苷酸对转录因子与DNA序列的结合能贡献是独立的，即所谓的可加性假设。Benos等的分析表明，虽然可加性假设并不总是成立，PWM模型并不完美，但它仍然是对TFBS结合能力的一个较好的近似。

近年来，有很多研究者从实验和计算的角度对"可加性假设"是否成立的问题进行讨论，他们的研究都表明，某些TFBS上的不同位置的核苷酸之间表现出明显的相关性。为了增强模型的预测能力，研究者尝试放宽可加性假设，提出包含相关性的模型来描述TFBS。Barash等提出用贝叶斯网络模型来描述TFBS，Zhou等将PWM加以扩展，提出了广义位置权重矩阵（generalized position weight matrix，GPWM），该模型考虑了互不重合的任意两个位置间的相关性，他们发现大约25%已知TFBS有比较强的位置相关性，而应用新模型GPWM后，大约有80%的TFBS预测识别率会有所提高，Gunewardena等提出MonoDi-Nucleotide模型，该模型假设TFBS上的任意一个核苷酸，或者是独立于其他核苷酸，单独贡献结合能，或者是与相邻的一个核苷酸相互作用共同贡献结合能，并采用动态规划算法进行优化，选出TFBS中有相互作用的相邻核苷酸对，Sharon等提出了"特征模块模型"（feature motif model，FMM）。FMM本质上说是一个对数线性模型（log-linear model），它假设一段DNA序列是TFBS的概率的对数与这段序列"特征"的加权和成正比。该模型中的"特征"指的是将TFBS序列映射为数值的函数，此特征可能与一个核苷酸有关，也可能多个核苷酸有关，因此能够描述不同位置上核苷酸间的相关性。特征定义的灵活性使得FMM模型可以描述任意多个位置上核苷酸间的相关性，为了避免引入与结合能无关的特征，在参数估计的同时进行变量选择，从而避免过度拟合。

相对于PWM，这些模型在不同程度上允许了TFBS不同位置上核苷酸间的相关性，但也对相关性有一定的限制，如贝叶斯网络有"无环假设"，GPWM仅考虑了任意两个位置上核苷酸间的相关性，而MonoDi-Nucleotide模型只考虑了相邻的两个位置上核苷酸间的相关性。虽然，从理论上说FMM可以描述任意的相关性，但在应用中仍受到样本量的限制，实际上作者也只考虑了可以由两个核苷酸决定的特征。事实上，在样本量有限的情况下，模型描述的相关性越复杂，它的表达能力就越强，但引入的参数也就越多，越容易造成过度拟合的现象，影响模型的稳健性和预测能力。上面提到的模型都在对PWM模型有所改进的情况下，在模型表达能力和稳健性之间作了不同程度的折中。

五、利用从头预测算法识别转录因子结合位点

de novo预测算法的基本逻辑是，以一组共调控的基因作为输入，用计算方法搜索在这些基因的上游调控序列中富集的motif。此类算法有很多，如基于EM算法的MEME，基于贪婪算法的Consensus，基于"词穷举法"（word enumeration）的Seeder，基于吉布斯抽样（Gibbs Sampler）的AlignACE、MotifSampler、BioProspector等等，详细的在线资源列表见表7-11。在多种软件并存的情况下，它们之间预测准确率的比较成为研究者关心的问题。然而，由于对转录调控过程、转录因子与DNA结合过程缺乏透彻的了解，缺乏标准数据，缺乏合适的评价标准，这个问题并不容易回答。2005年，Tompa等对13种de novo

预测软件进行了系统的评测。他们从TRANSFAC中提取TFBS信息构建正负样本集，应用这些软件做TFBS预测，并提出多种指标(从单个核苷酸，TFBS整体两个不同的层次衡量预测算法的敏感度等)来评测算法的表现。他们的分析表明：各种软件之间没有绝对的优劣，软件的绝对检测效果都不是太高。在Tompa的标准下(如果预测出的TFBS与真TFBS有重叠，并且重叠的长度超过真TFBS长度的1/4，就认为预测是准确的)，13个软件中最高的灵敏度(Sensitivity)为0.22。另外，不同软件的预测效果对不同的数据集、不同的物种有明显的偏好性，而且大部分软件在酵母数据上的效果明显高于其他物种，这与TRANSFAC中酵母数据的相对丰富是分不开的，如果允许软件同时预测出两个motif，预测的准确率有可能得到提高。最近Wijaya等开发的MotifVoter通过综合不同预测算法的结果进行预测。在Tompa等构造的测试集上，MotifVoter的敏感度比单个的预测算法提高了275%。

表7-11　de novo预测TFBS的软件

软件名称	网址
MEME	http://meme.sdsc.edu/
Consensus	http://bifrost.wustl.edu/consensus
Seeder	http://www.cpan.org
AlignACE	http://atlas.med.harvard.edu/
MotifSampler	http://www.esat.kuleuven.ac.be/~dna/Biol/Software.html
BioProspector	http://ai.stanford.edu/~xsliu/BioProsputor/
MotifVoter	http://www.comp.nus.edu.sg/~bioinfo/MotifVoter

另外，考虑到转录过程是由多个转录因子组合调控的，Zhou等提出了CisModule算法来预测多个TFBS构成的模块。在模拟数据集以及真实数据上，CisModule都能准确的预测出TFBS模块，而且对单独的TFBS的敏感度也优于普通的de novo预测算法(在与果蝇早期发育相关的基因构成的数据集上，CisModule灵敏度达到56%，而MEME在相同的数据集上的灵敏度仅为9%)，这说明利用多个转录因子的合作信息能够提高预测的准确性。关于整合组合调控信息预测TFBS的相关算法，Hannenhalli在最近的一篇综述中有更详细的介绍。

de novo预测算法有局限性，它依赖于预先构建的共调控的基因集合。这个基因集合的构建通常来自于基因功能的分析，比如生物芯片的表达数据，ChIP-chip实验等。在很多情况下，这些功能信息是不易获得的；另外，对“共调控”信息的依赖，也使得de novo检测算法局限于对单物种的分析。

六、结合Chip-seq等高通量实验数据的转录因子结合位点预测方法

随着基因芯片等高通量数据的出现，计算方法在转录因子结合位点的分析中得到了广

泛的应用。对转录因子结合位点的计算研究可分为两类问题：第一类问题是通过收集可能被同一转录因子调控的基因启动子序列，在其中寻找具有统计显著性的短片段，作为转录因子可能的结合位点；第二类问题是根据若干已知的转录因子结合位点的模体，在所研究基因的启动子区域内搜索相应转录因子可能的结合位点。

一张基因芯片（microarray）可以同时检测数万个基因在某个组织样本中的表达值，对在不同条件下获得的基因芯片数据进行聚类分析，我们可以得到一组或几组有相似表达模式的基因。它们在特定的组织，或者特定发育阶段被同时激活或同时抑制。由此推断，这些基因很可能受到共同转录因子的调控。相同的转录因子在这些基因启动子区域上的结合位点应当是相同或者相似的。通过计算方法寻找这些相似的转录因子结合位点（模体），称为转录因子结合位点的识别。把输入启动子序列看作一些杂乱无章的背景噪声，模体可以看作隐藏在背景噪声中的有规律的信号。通过计算方法，我们希望找到那些出现次数明显高于其他背景噪声的信号。这里我们需要注意两点，第一是我们需要一组可能含有共同调控元件的序列，从中发现某种频繁出现的“信号”，不可能只从一个序列中找到模体；第二是输入序列中的“信号”要足够强，可以同背景噪声区分。这样一组共调控的基因除了可以通过对基因芯片数据进行分析得到，也可以通过对已有知识进行总结得到。比如处于同一个通路（pathway）上的功能相关的基因也可能被同一转录因子调控。找到一组共调控的基因之后，首先遇到的一个问题就是如何确定基因的启动子区。一般认为，转录因子结合位点主要在转录起始位点（transcription start sites，TSSs）附近出现，但还有一些转录因子结合在基因上游很远的区域（被称为远程作用）。根据研究问题的不同，启动子序列的长度可以取几百到几千个碱基不等，通常选取转录起始位点附近1000~2000个碱基的长度作为启动子区（例如，转录起始位点上游1000 和下游200 个碱基）。序列太短会丢失部分结合位点。如果序列取的过长，在包含了少量真实结合位点的同时，却引入了大量的背景噪声，使真正的转录因子结合位点淹没在噪声中无法区分。近年来，ChIP-chip 和ChIP-seq 技术在转录因子结合位点的分析中得到了广泛应用。与由基因芯片和功能相关获得的包含共同转录因子结合位点的启动子序列相比，ChIP-chip 和ChIP-seq 确定的包含共同结合位点的区域更加准确。得到一组含有共同结合位点的候选启动子序列后，就可以利用已有的计算方法进行结合位点的识别，然后对结果进行后续处理并解释它们的生物意义。在这里候选序列集合起到了“训练集”的作用，它的选取对后续分析结果的影响非常大，应尽量选择包含信号可能性大的序列，序列的数目以数十到数百条为宜，如果序列过少，可以考虑加入较近的直系同源序列。图7-12所示为转录结合位点识别的完整流程。

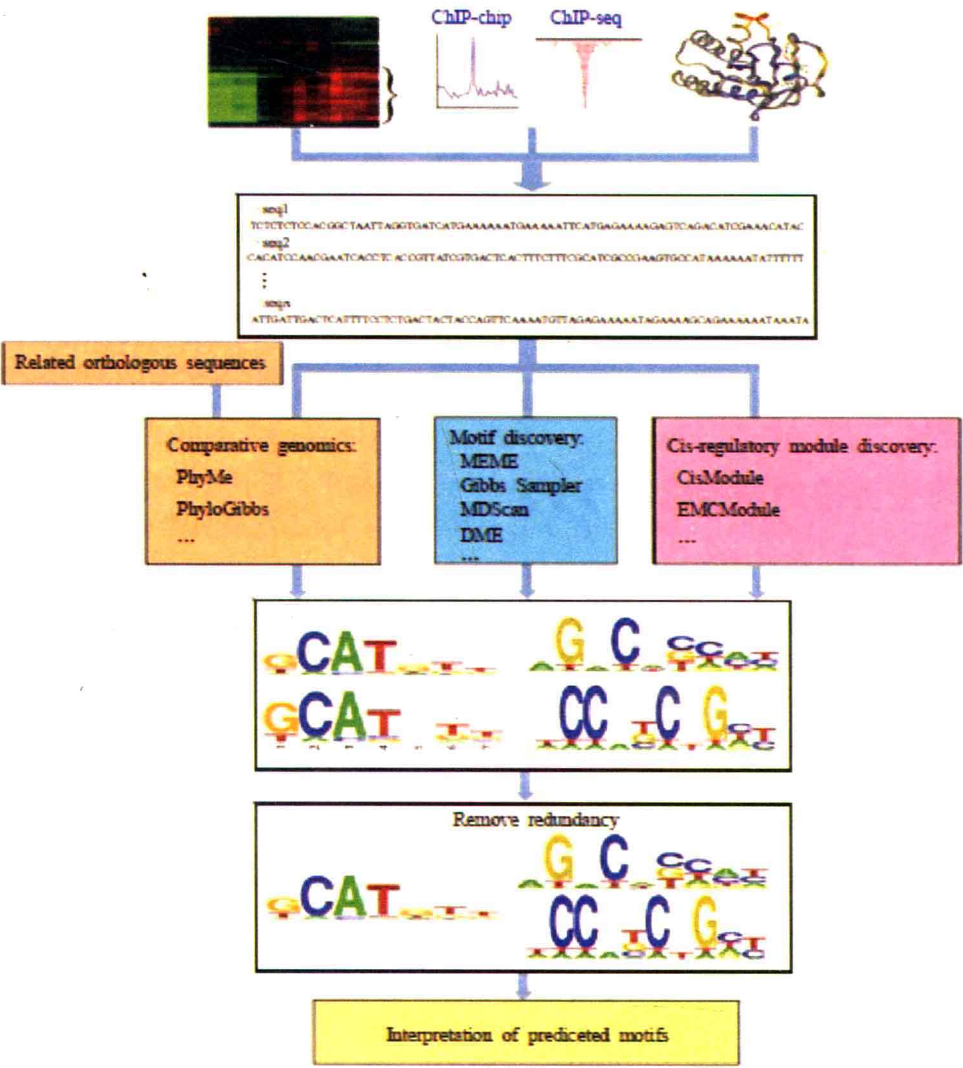


图7-12　模体(motif)基本分析流程示意图

来源自LI Ting-ting, JIANG Bo, WANG Xiao-wo, et.al., *tutorial for computational analysis of transcription factor binding sites*, acta biophysica sinica, 2008.24：334-347.

第四节

可变剪接的生物信息学研究

Section 4 Bioinformatics Analysis on Alternative Splicing

一、可变剪接的调控机制

可变剪接是指从一个mRNA前体中通过不同的剪接方式(选择不同的剪接位点组合)产生不同的mRNA剪接异构体的过程。可变剪接是调节基因表达和产生蛋白质组多样性的重要机制。剪接过程受多种顺式作用序列和反式作用因子相互作用调节。包括SR和hnRNP家族蛋白在内的多种剪接因子参与这一调节过程。转录机器(machine)也参与可变剪接的调节。

(一) 可变剪接与蛋白质组多样性

据预测,人类基因组可能有约35 000个基因,果蝇约14 000个,而简单的模式生物线虫约19 000个基因。生物的复杂性与其基因组基因数量似乎存在明显差异,原因在蛋白质组。基因重排、RNA编辑和可变剪接等机制可以从一个基因产生多种蛋白,从而使蛋白质组中蛋白质的数量超过基因组中基因的数量。其中,从影响的基因数量和生物种类范围来看,可变剪接是扩大蛋白质多样性的最重要的机制。

(二) 可变剪接的频率

1. 5%。从1977年Walter Gilbert提出可变剪接概念,1980年Baltimore在小鼠IgM基因发现第一个可变剪接产生膜型、分泌型IgM,至2001年,用经典分子生物学实验的方法研究,一共仅发现了数百种有可变剪接的基因。并推测在高级真核细胞生物约5%的基因有可变剪接。

2. 35%~60%。高通量的基因组测序和EST测序,使得生物信息学的方法研究可变剪接成为可能。EST来源于完全加工的mRNA,它们提供了一个广泛的mRNA多样性的样品库。这种多样性可以用计算机分析。最近两年,多个研究小组通过不同的生物信息学的方法,从整个人基因组的水平进行分析,结果一致显示约35%~60%的人基因有可变剪接形式。而且,由于对大多数基因来说,每个基因只测到了很少几个EST甚至没有EST; EST不是全长的mRNA,多位于mRNA的5’和3’端; EST来源于有限的组织和发育阶段; 很有可能存在有更多的可变剪接而在现在的EST库中没有显示。因此实际可变剪接的频率可能比预测的更高。这还有待于建立新的高通量的分子生物学方法,如生物芯片的方法,以进一步实验验证。

(三) 单个基因可变剪接产生的多样性

一个基因可以通过如下几种方式产生多个转录体,如不同的转录起始位点,可变剪接,选择不同的加尾信号位点,RNA编辑等。可变剪接包括3种类型:①内含子的保留;②可变外显子的保留或切除;③3'和5'剪接位点的转移(shift)导致外显子的增长或缩短。可变剪接对蛋白质结构的影响也是多样性的,如多肽链中一个到数百个氨基酸的增加或减少;某功能域的有无;如果可变剪接使读码框架改变,则可能无法有效翻译,mRNA被监视系统降解。

单独一个基因通过可变剪接产生的十几种剪接异构体的现象很常见。有些基因甚至能够产生成千上万种剪接异构体。最突出的例子是果蝇(Drosophila melanogaster)的*Dscam*基因,可以通过可变剪接产生38 000多种mRNA异构体。*Dscam*基因编码一个神经元轴突定向受体,它细胞外有一个由10个免疫球蛋白重复序列组成的结构域,第2,3,7个免疫球蛋白重复序列分别由第4,6,9号外显子编码,4号外显子盒(cassette)有12个变异体,6号外显子有48个变异体,9号外显子有33个变异体,再加上17号外显子的2个变异体。每个成熟的Dscam mRNA分别只有一个有4,6,9,17号外显子的变异体,由此理论推测*Dscam* 基因共有 $12 \times 48 \times 33 \times 2 = 38\ 016$种剪接异构体。对*Dscam*基因50个cDNA克隆随机测序发现了49种不同的剪接异构体,说明实际存在的剪接异构体即使没有理论那么多,也至少有上千种。人的*Neurexins*, *n-Cadherins*, calcium-activated potassium channels等基因也有类似的高度多样的剪接异构体。

上述现象非常类似于淋巴细胞TCR或免疫球蛋白的胚系基因重排,不同之处在于后者发生在DNA水平,前者发生在RNA水平。基因重排产生的高度多样抗原受体库可以识别高度复杂的自身和异己抗原。而*Dscam*基因的转录异构体可能有神经系统的发育有关。神经元的定向迁移和相互连接可能是发育过程中最复杂的事件。果蝇约有25 000个神经元,要使它们生长的轴突准确地、可重复地到达目的,使这些神经元准确地连接在一起,必然需要一个特殊的系统。*Dscam*基因的38 000多种mRNA异构体,每个异构体各编码一个不同的受体,每个受体具有识别不同分子定向信号的潜能,从而有能力指导各个生长的轴突到达准确的位置。

如果将可变剪接与其他RNA加工过程(如RNA编辑)联系起来共同考虑,基因产物会更复杂。例如,果蝇的*para*基因(voltage-gated action potential sodium channel)有13个可变外显子,可编码1536种不同的mRNA,另外,*para*的转录体还要经过在11个已知位点的RNA编辑,这样理论上一共可以产生1 032 192个不同的*para*转录异构体。

根据受可变剪接影响的基因的概率,以及单个基因可能产生的可变剪接体的数目,足以表明可变剪接对蛋白质组多样性的巨大影响。

(四) 可变剪接的功能和生物学意义

1. 可变剪接是在RNA水平调控基因表达的机制之一　一个基因通过可变剪接产生多个转录异构体,各个不同的转录异构体编码结构和功能不同的蛋白质,它们分别在细胞/个体分化发育不同阶段,在不同的组织,有各自特异的表达和功能。因此,可变剪接是一种在转录后RNA水平调控基因表达的重要机制。

目前已知的可变剪接异构体中,只有一小部分明确确定了功能和生物学意义。第一个确定的可变剪接异构体功能是 IgM基因,其末端最后两个外显子的可变剪接,决定了所编码

的膜型/分泌型IgM的产生。最著名的例子是果蝇性别决定系统,在此系统中,至少5个基因(*sxl*, *tra*, *msl2*, *dsx*, *fru*)转录体的可变剪接级联反应最终决定了果蝇雄性和雌性性别特征的表达。有些基因,可变剪接造成的蛋白质异构体之间功能上的差异没有被实验检测出来。不过阴性的结果不能代表没有功能差异,只是目前没有检测出来而已。也有很多异构体造成读码框架改变,不能被翻译为蛋白质,而是直接被降解了。真核生物也有mRNA监视系统NMD(nonsense-mediated degradation),检测 mRNA中异常提前出现的终止密码子,一经发现,立即降解异常的mRNA,防止其翻译。在大多数情况下,检测可变剪接造成的蛋白质异构体之间功能上的差异的实验还没有开展。最近发展的RNAi技术,可以适应高通量的从功能基因组水平研究各基因可变剪接异构体的功能的要求。2000年已经有人将RNAi技术应用于模式生物线虫的可变剪接异构体的大规模研究上(目前已经大量开始用于哺乳动物系统)。

2. 多样性与复杂性　可变剪接是从相对简单的基因组提高蛋白质组多样性的重要机制,蛋白质组的多样性与多细胞高等生物的复杂性相适应。从可变剪接涉及的基因分布格局分析,可变剪接多发生在参与信号传导和表达调节等复杂过程的基因上,如受体、信号传导通路(凋亡)、转录因子等。对个体分化发育和一些关键的细胞生理过程如凋亡、细胞兴奋等的精确调控有重要意义。从可变剪接涉及的基因系统分类分析,可变剪接多发生在免疫和神经等复杂系统。正如*Dscam*基因所示,可变剪接产生的多样性,赋予这些系统精确处理复杂信息相适应的潜力。

(五) 可变剪接的调节机制(图7-13和图7-14)

可变剪接能够产生惊人的多样性,但我们对其调节机制所知不多。剪接位点的选择受到结合到非剪接位点RNA元件的剪接因子的多重调节。参与可变剪接调节的RNA元件包括ESE、ISE、ESS、ISS。剪接因子包括SR和hnRNP家族蛋白等多种因子。

真核生物新生的mRNA前体经过5’戴帽,剪接,3’加尾等加工成为成熟的mRNA。在剪接反应过程中,含有内含子和外显子的新生的mRNA前体,在剪接体作用下切除内含子,并将外显子依次连接起来的过程。剪接反应由剪接体执行,剪接体包括5个小核糖核蛋白复合体U1、U2、U4、U5 和U6 snRNPs,和50~100种非snRNP蛋白。剪接体通过RNA-RNA, RNA-蛋白质,蛋白质-蛋白质等多重相互作用以精确切除每个内含子和以正确次序连接外显子。

为有效剪接,绝大部分内含子需要:

1. 一个保守的5’剪接位点,A/CAG↓GURAGU。
2. 一个分支点序列BPS ,YNYURAY,后面跟着一个多聚嘧啶Pytract Y10-20。
3. 一个3’剪接位点YAG。

剪接体的形成是一个多步骤依次进行过程,形成多个中间体:

1. E -复合体形成　U1snRNA通过碱基互补识别5’剪接位点,SR蛋白结合。U2AF65和U2AF35识别多聚嘧啶Pytract和3’剪接位点。
2. A -复合体形成　U2snRNA通过碱基互补识别分支点序列BPS; 需ATP。
3. B -复合体形成　U4/U6 _ U5 tri-snRNP随后与mRNA结合。
4. C -复合体形成　最后, RNA-RNA, RNA-蛋白质相互作用构象改变形成有催化活性的剪接体。

发现新的可变剪接异构体,确定每个异构体的独特功能和生物学意义,并阐明其调节机

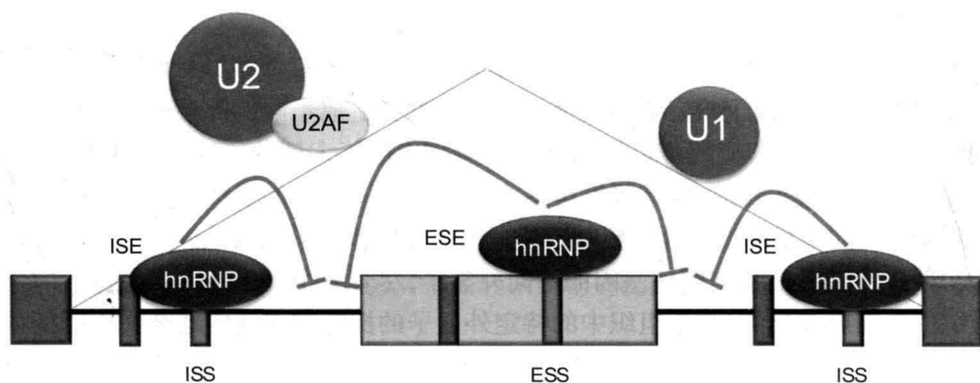


图7-13 剪切的抑制

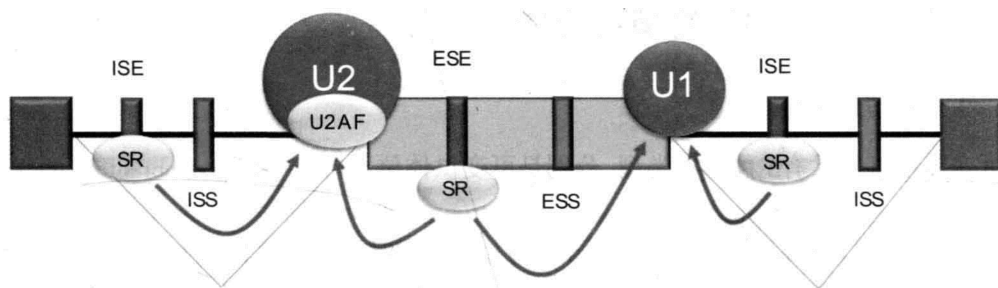


图7-14 剪切的激活

制，是功能基因组时代研究的一个重要领域。在这一领域研究中，除利用经典的分子生物学技术外，还需建立新的高通量的技术，如生物芯片技术，RNAi技术等，并要与生物信息学技术紧密结合，同时需要细胞生物学、生物化学、临床与病理学、免疫学等多学科的协作，才有可能对这一重要的生命现象有所了解。

二、可变剪接数据资源：ASD、ASTD

可变剪接是真核生物有别于原核生物的基本特征之一，近年来随着大量测序工作的开展，通过实验和计算机处理的方法已经确定了越来越多的可变剪接事件，研究人员也建立了很多与可变剪接相关的数据库。例如，ASAP(alternative splicing annotation): http: //www.bioinformatics.ucla.edu/ASAP，AS-ALPS(alternative splicing-induced alteration of protein structure): http: //as-alps.nagahama-i-bio.ac.ip，ASTD(alternative splicing and transcript diversity): http: //www.ebi.ac.uk/astd/等数据库。ASD(altemative splicingdatabase)数据库，网址是: http: //www.ebi.ac.uk/asd/，现在ASD与ATD数据库合并成数据库ASTD。

ASD数据库包括人类和老鼠等多种模式生物的可变剪接事件和剪接异构体，提供了AltExtron、Altsplice以及AEdb三个子库。

三、利用基因芯片技术进行可变剪接研究

大量确定选择性剪接的实验数据需要用生物信息学的方法来分析，其中最有效的方法

就是用较长探针（大于60nt）的微阵列来分析。Schoemaker等人用这种技术检测了在人类22q染色体上注释的8183个外显子。这种技术很适合去检测选择性剪接，设计一条跨越外显子-外显子连接处的探针，由于给定基因选择性剪接产物会造成不同的外显子-外显子连接处，与不同组织mRNA样品杂交就可以检测选择性剪接。尽管大部分给定基因的外显子-外显子连接处的杂交率是不变的，但选择性剪接会导致一些连接处的上移或下移。这些芯片的快速出现使从不同组织的选择性剪接基因编写出选择性剪接形式目录成为可能。Affymetrix公司用20种探针（25nt）代表同一基因的不同外显子，尽管一个基因的大部分探针的强度在不同组织中会有所变化，但某个组织中的特定外显子的探针被不规则地杂交可以指示选择性剪接。但要指出的是只用基因芯片的方法，不结合生物信息学分析，是无法解决选择性剪接识别的问题。

四、RNA-seq与可变剪接异构体

(一) 可变剪接事件

可变剪接事件共有5 种基本类型，分别是可变供体位点（alternative donor site）、可变受体位点（alternative acceptor site）、内含子保留型（intron retention）、外显子缺失型（exon skipping）和外显子互斥型（mutually exclusive exon）。另外也有分为7 种形式的，包括前面5 种类型加上可变的起始或末端外显子，而后两种形式更有可能是可变启动子和可变polyA 位点造成的，可进行专门分析。

绝大多数真核基因编码序列由外显子和内含子间隔组成。外显子和内含子之间的边界称作剪接位点，按它们在内含子两端的位置又可分为5′ 剪接位点（位于内含子的5′ 端，也称作供体位点）和3′ 剪接位点（位于内含子的3′ 端，也称作受体位点）。基因的前体mRNA 被转录后，必须通过剪接反应切除内含子，把外显子连在一起，形成一个成熟的mRNA，由细胞核转运到细胞质中进行翻译。可变剪接（alternative splicing）是指从一个mRNA前体中通过不同的剪接方式（选择不同的剪接位点组合）产生不同的mRNA 剪接异构体，生成具有不同化学性质和生物功能的蛋白亚型的过程。可变剪接是高等真核生物中丰富蛋白质多样性的重要机制之一，非正常的可变剪接会导致各种疾病。

(二) RNA-seq与可变剪接异构体

剪接位点的精确定位是确定真核生物基因结构的关键，目前有多种方法可用来在基因组范围内识别剪接位点。RNA-seq技术是全新的转录组研究方法，基本上克服了上述技术的弊端和缺陷，无需预先设计探针，可对任意物种的整体转录活动进行检测，发现新基因、新剪接位点和可变剪接事件，对转录体结构的分析有了明显的提高。

RNA-Seq 还可对可变剪接（alternative splicing）进行定量研究。Sultan 等利用深度测序对人类细胞系mRNA 剪接进行了全局性研究，鉴定出94 241 个剪接位点，其中有4096个是全新的。该研究还表明，外显子跳跃（exon skipping）是选择性剪接的一种普遍形式。最新RNA-Seq 数据分析显示，至少48%的水稻基因经历可变剪接，比之前报道的利用RNA-Seq数据分析结果（33%）和EST/cDNA 数据分析结果（20%~30%）多；在拟南芥中，至少42%携带内含子的基因经历可变剪接，多于之前利用EST/cDNA 数据分析的20%到30%，并且这些

可变剪接转录本中，大多数是携带成熟前终止密码子的剪接异构体，可能在基因表达调控中发挥重要作用。

根据RNA-seq 技术的最新应用（如图7-15：利用RNA-seq数据重构转录本），人们越来越多地发现即使来自同一基因的剪接异构体也可能具有不同的功能。因此，传统的根据结构基因组学将基因定义为“基因组上可定位的一段区域、可被遗传的基本单元”面临着巨大的挑战。而根据这种基因定义构建的功能注释数据库也将面临较大的改进。随着第三代单分子测序技术的发展，我们将有机会对基因转录产物进行更深入细致的研究，其带来的不仅是技术的革新，更是知识的革新。

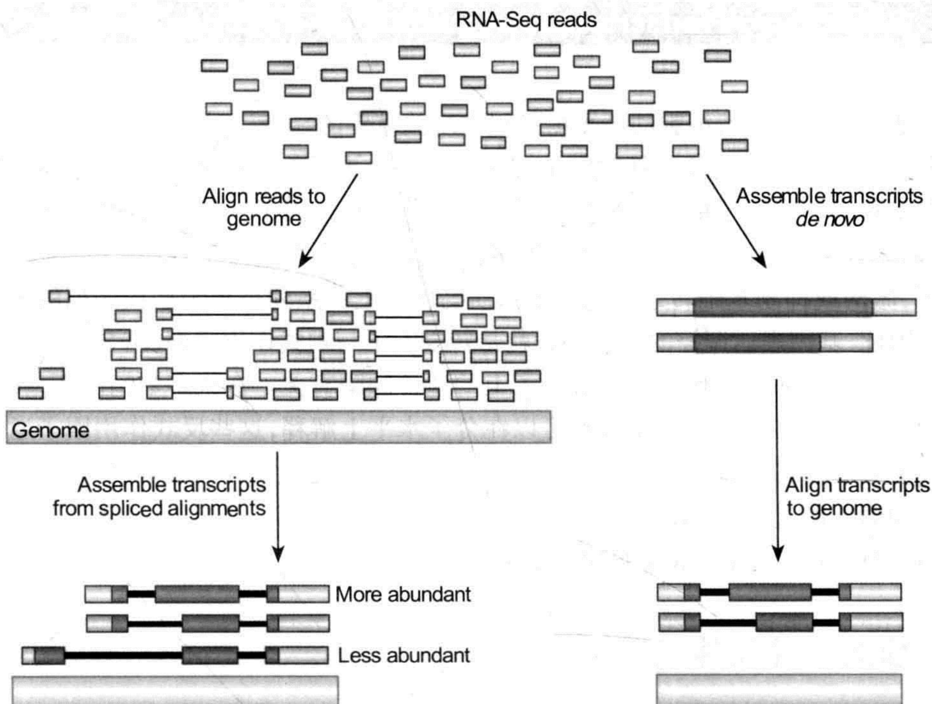


图7-15 用RNA-seq数据重构转录本的策略

来源自：1. Trapnell C, Pachter L, Salzberg SL. , *TopHat: discovering splice junctions with* RNA-Seq. Bioinformatics,2009. 25(9): 1105-1111. 2. Haas BJ, Zody MC., *Advancing RNA-Seq analysis*, Nature Biotechnology,2010. 28 : 421-423.

第五节

转录调控与人类疾病

Section 5 Transcription Regulation and Human Disease

基因表达调控是维持细胞动态平衡过程中至关重要的一步。对基因表达的控制可以有多个步骤。绝大多数的调控事件都发生在转录水平。为了起始转录,真核生物RNA聚合酶Ⅱ需要与被称为转录因子的一系列蛋白质密切合作。转录因子一般分为两组: ①基础转录因子,它们无处不在,并且招募RNA聚合酶Ⅱ多重蛋白复合物到最小启动子; ②基因特异的转录因子,激活或抑制基础转录。这些蛋白质与DNA上一系列被称为调控序列的调控模块相结合。因此,基因表达调控的分子基础就是转录因子与顺式作用序列(结合位点)的结合。越来越多的研究表明人类疾病与转录因子上的遗传缺陷有关。在大多数情况下,转录因子处的基因突变导致多效性。临床观察可以在分子水平上解释,这些反式作用因子通常通过与一个或更多的深层激活子结合,从而控制许多基因的表达。此外,许多事件导致白血病和实体瘤的肿瘤起源过程,暗示着转录因子的过表达或基因突变。这一节描述了归因于转录因子编码基因及其同源结合位点处突变造成的人类疾病。

一、顺式调控元件

转录起始复合物与RNA聚合酶Ⅱ和其他相关的基础因子(一般转录因子)共同装配,进而起始转录。这种多蛋白复合物与一段被称为核心启动子的短DNA序列结合,这段短DNA序列往往包含一个位于转录起始位点上游20~30个碱基,被称为TATA盒的保守模序(motif)。一般转录因子的特点是它们能够控制最小启动子上RNA聚合酶Ⅱ的活性。这一步是它能够有效起始转录的关键,但是它的调控还需要与不同调控元件结合的其他因子的介导。

下面将简要介绍这些调控元件:

顺式作用元件(cis-acting element)存在于基因旁侧序列中能影响基因表达的序列。顺式作用元件包括启动子、增强子、调控序列和可诱导元件等,它们的作用是参与基因表达的调控。顺式作用元件本身不编码任何蛋白质,仅仅提供一个作用位点,要与反式作用因子相互作用而起作用。

顺式作用元件是指与结构基因串联的特定DNA序列,是转录因子的结合位点,它们通过与转录因子结合而调控基因转录的精确起始和转录效率。

顺式作用元件是转录调节因子的结合位点,包括启动子、增强子和沉默子。真核基因

启动子是原核启动序列的同义语。真核启动子是指RNA聚合酶及转录起始点周围的一组转录控制组件，每个启动子包括至少一个转录起始点以及一个以上的功能组件，转录调节因子即通过这些功能组件对转录起始发挥作用。在这些调节组件中最具典型意义的就是TATA盒子，它的共有序列是TATAAA。TATA盒子通常位于转录起始点上游-25至-30区域，控制转录的准确性和频率。TATA盒子是基本转录因子TFⅡD结合位点；TFⅡD则是RNA聚合酶结合DNA必不可少的。除TATA盒子外，GC盒子（GGGCGG）和CAAT盒子（GCCAAT）也是很多基因中常见的，它们位于起始点上游-30至-110bp区域。所谓增强子就是远离转录起始点、决定组织特异性表达、增强启动子转录活性的特异DNA序列，其发挥作用的方式与方向、距离无关。增强子与启动子非常相似：都是由若干组件组成，有些组件既可在增强子、又可在启动子出现。从功能方面讲，没有增强子存在，启动子通常不能表现活性；没有启动子，增强子也无法发挥作用。增强子和启动子有时分隔很远，有时连续或交错覆盖。

某些基因有负性调节元件抑制子（沉默子）存在。有些DNA序列既可作为正性、又可作为负性调节元件发挥顺式调节作用，这取决于不同类型细胞中DNA结合因子的性质。

核心启动子上游的顺式作用序列，以一个依赖方向的方式，发现了所谓的近启动子元件；这些序列被特定的转录因子绑定，它们的出现可以增加或减少基因的转录活性。除了启动子区域，其他顺式作用元件可以位于起始位点5’或3’端几百或上千碱基对内。这些元件也是序列特异的转录因子的结合位点。与启动子相比，这些元件的位置和方向是关于基因的变量。如果特异因子与这些元件的结合可以激活转录，那么这些元件被称为增强子；如果抑制转录，则称为沉默子。由于与这些元件结合的转录因子可能在不同环境、不同组织中具有不同的功能，导致特定的顺式作用元件的重要性在不同的细胞类型和对不同生理刺激的反应上有很大的区别。

多因子重叠或叠加的结合位点可以导致不同的阳性和阴性因子对位点的竞争。在某些情况下，协同效应依赖于顺式作用元件附近严格的间距。各种类型的沉默子元件可以阻断顺式连接增强子的活性。

二、反式作用因子

反式作用因子（trans-acting factor）是指能直接或间接地识别或结合在各类顺式作用元件核心序列上参与调控靶基因转录效率的蛋白质。

大多数真核转录调节因子由某一基因表达后，可通过另一基因的特异的顺式作用元件相互作用，从而激活另一基因的转录。这种调节蛋白称反式作用因子。

参与基因表达调控的因子，它们与特异的靶基因的顺式元件结合起作用。编码反式作用因子的基因与被反式作用因子调控的靶序列（基因）不在同一染色体上。反式作用因子有两个重要的功能结构域：DNA结合结构域和转录活化结构域，它们是其发挥转录调控功能的必需结构，此外还包含有连接区。反式作用因子可被诱导合成，其活性也受多种因素的调节。

同一类序列特异性的反式作用因子由多基因家族所编码，它们具有特定的蛋白质结构（如上述的锌指结构、碱性亮氨酸拉链、螺旋-环-螺旋基元等）和蛋白质结构上的同源性，因而构成反式作用因子家族，如类固醇激素受体家族、AP1家族等。

主要包括:

1. DNA结合域 ①螺旋-转角-螺旋; ②锌指结构; ③亮氨酸拉链; ④螺旋-突环-螺旋。

2. 转录激活域 与其他转录因子相互作用的结构成分。

随着表观遗传学的发展,研究发现除了蛋白, DNA、RNA也有调控功能,所以现在也称反式调控元件,主要有miRNA、转录因子等。

三、顺式作用元件与疾病(心脏病、肾病、Alzheimer、cancer)

(一) B Leyden血友病

X-连锁因子Ⅸ基因中的突变可以导致乙型血友病。多数情况下是由于编码蛋白质序列中的突变。然而,在少数情况下,疾病归因于因子Ⅸ基因调控区域中的突变。B Leyden血友病患者伴有严重的出血症状并且<1%在童年时血浆凝血因子Ⅸ量正常。青春期后,临床症状逐步改善并且血浆凝血因子Ⅸ浓度上升到正常人的60%。所有被研究的患者在因子Ⅸ基因转录起始位点附近20bp范围内存在突变。这些突变扰乱了转录因子与因子Ⅸ基因的结合,这对因子Ⅸ基因表达来说是至关重要的。例如,在-20处的突变干扰了肝细胞核因子4(HFN4)的结合。此外,因子 Ⅸ的启动子-22到-38区域包含一个雄性激素受体结合位点,这一位点与HFN4结合位点有交叠。在青春期,雄性激素受体与这一位点结合可以补偿缺乏HFN4或其他转录因子,激活因子Ⅸ基因。某些-22到-38区域的突变,被称为Brandbourg变异,可以阻止这种补偿,导致在青春期没有任何改善。

(二) 血红蛋白病

遗传性持续性胎儿血红蛋白增多症(hereditary persistence of fetal hemoglobin, HPFH), g-球蛋白在成年后仍然持续表达,可以作为另一个由顺式作用元件突变导致人类疾病的例子。在Ag-球蛋白基因启动子区域已经识别出了点突变,在那里存在GATA-1转录因子结合位点。Ag-球蛋白基因不能被抑制。GATA-1结合位点也在LCR中存在。这可以部分解释在西班牙裔地中海贫血中到底发生了什么,大部分LCR叶基因簇在染色质构象中删除,导致DNA酶 Ⅰ不可接近,造成球蛋白基因表达缺乏。TATA盒(-28到-31)和CACC盒(-92到-105)的突变已经在b-地中海贫血中发现,其特点是b-球蛋白基因表达减少。

(三) 进行性肌阵挛性癫痫

Unverricht-Lundborg type进行性肌阵挛性癫痫是一种罕见的常染色体隐性遗传疾病,发病6至13年带有不同程度的精神恶化及小脑共济失调。

(四) 启动子多态性与人类疾病

影响某些基因表达水平的启动子多态性可能与各种不同病症有关。不同的MHC Ⅱ类等位基因的差异表达可能与启动子的多态性有关。例如,一些HLA-DQ基因的等位基因具有不同强度的启动子,以响应细胞因子,如TNF-α,揭示出与某些自身免疫疾病易感性等位基因相关的致病机制。人类TNF-α启动子的一个稀有等为被称为TNF2,位于一个被明确定义的与自身免疫和高TNF-α(肿瘤坏死因子)产生相关的HLA-A1单体附近。TNF-α的高血

浆水平与疟疾和利什曼原虫感染的严重性相关。此外，为TNF2等位基因纯合子的患者表现出脑型疟疾死亡率明显更高。

四、反式作用因子与疾病

转录因子的重要作用以及单一因子可以影响许多基因表达这一事实表明由遗传突变导致的转录因子的失活与生存是对立的。这在许多情况下可能是正确的，但一些转录因子的突变与生存兼容，并且导致了特定的疾病。在过去几年，越来越多地发生在编码基因转录因子处的基因突变表现出与一系列先天性综合征相关。其后果是畸形、生理通路的中断或者肿瘤发生。观察到的异常经常局限于表达受影响基因组织的子集。在大多数情况下，表型是多效的，反映了转录因子控制许多基因表达这一事实。突变分析，与同源小鼠模型比较，揭示出蛋白质共同作用的分子机制，深入观察了由这些基因控制的主要生理过程。

除了先天综合征，大量特定转录因子的体细胞突变促成了肿瘤发生的多步骤过程并且导致越来越多的癌症，在这些癌症中，这些步骤可能会起到一定的作用。在某些肿瘤中，如人类白血病，观察到的染色体易位，就是多种转录因子基因的调控和编码区域重排的结果（图7-16）。

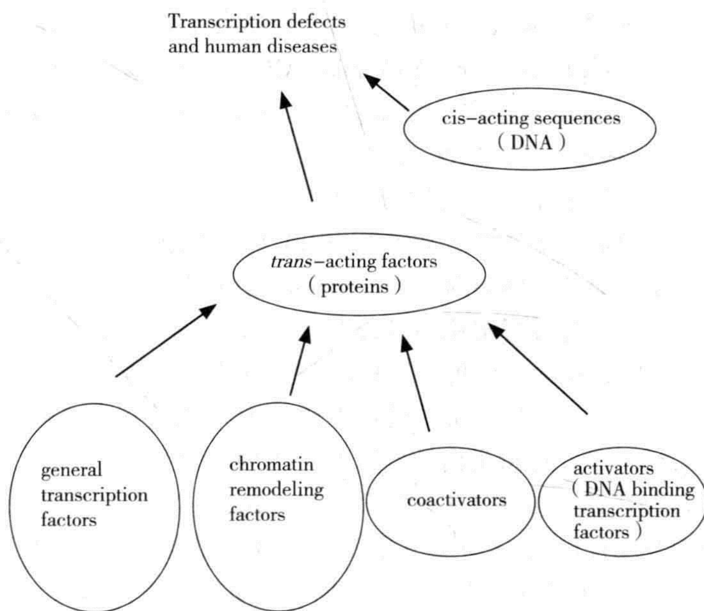


图7-16 转录调控缺陷与人类疾病

来源自：Jean Villard, *Transcription regulation and human diseases*, SWISS MED WKLY, 2004.134：571-579.

（一）一般转录因子突变与人类疾病

在大量与由RNA聚合酶Ⅱ介导的转录起始阶段相关的一般转录因子中，TFIIH具有特殊的作用。这种多亚基蛋白复合物在一些受着色性干皮病困扰的患者中是缺乏的。着色性干皮病（xeroderma pigmentosum，XP）的特点是对阳光引起的皮肤伤害具有极度的敏感性，不

足或过度色素沉着，并易患皮肤癌。此病是由于缺乏一种与DNA修复相关的机制。TFIIH的两个亚基（XP-B和XP-D）对DNA核苷酸切除修复来说是必不可少的。这两种蛋白质都具有解旋酶活性。这两个亚基在某些XP患者中发生突变。此外，XP患者的一个亚基显示出眼部或神经系统异常，如精神迟缓以及身体和性征发育迟缓，但并不容易发生癌症。这些临床症状像DNA修复缺陷一样很难被合理的解释。患者携带XP-B或XP-D基因的突变具有XP和CS或TTD的临床特点。TFIIH在DNA修复和转录中发挥作用这一发现，引发了一种假设，即由编码TFIIH亚基的基因突变携带患者展示出的临床症状的非正常变异，并不是DNA修复缺陷的结果，而是来自于转录的缺陷。

（二）染色质重塑因子多态性与人类疾病中的共激活子

染色质重塑是人类表观遗传的重要方面，因此任何过程发生异常都会导致人类基因组的不正常表达，从而引起许多疾病。这其中，染色质重塑异常引发的人类疾病基本是由于重塑复合物中的关键蛋白发生突变，导致染色质重塑失败，即核小体不能正确定位，并使修复DNA损伤的复合物，基础转录装置等不能接近DNA，从而影响基因的正常表达而引起的。如果突变导致抑癌基因或调节细胞周期的蛋白出现异常将导致癌症的发生。乙酰化酶的突变导致正常基因不能表达，去乙酰化酶的突变或一些和去乙酰化酶相关的蛋白的突变使去乙酰化酶错误募集将引发肿瘤等疾病。

目前的研究中发现，白血病的发病机制中，染色质重塑异常是非常重要的一环。急性早幼粒细胞白血病（acute promyelocytic leukemia, APL）会导致多种染色体异常，结果形成PML2RARα、PLZF2RARα融合蛋白。然而在生理浓度的RA 存在时，PML2RARα并非激活转录而是阻抑转录，这是由于PML2RARα和N2CoR/ Sin3/ HDAC1 辅助抑制因子复合物间相互作用增强所致。当配基水平足以释放与野生型RARα结合的辅助阻抑复合物时，PML2RARα仍然和辅助阻抑复合物牢固结合，使RA 反应基因的启动子维持去乙酰化构象、阻抑转录，产生与RARα显性负抑制剂作用后相同的表型。

反式作用蛋白质可能通过影响染色质结构影响基因表达。在酵母中，有一组被称为SWI/SNF的基因，显示出很强的编码可以直接改变染色质结构的蛋白质的能力。这些蛋白作用的确切模式，即形成一个大的多蛋白复合物，目前还不能清楚的了解。这些蛋白质具有一个假想的DNA结合结构域（锌指蛋白结构域）如ATP酶/解旋酶类似的结构域。在人类中，SWI/SNF的几个同源基因已经被描述出来。其中之一是ATRX，它在X-连锁人类综合征中发生突变，可以导致神经发育迟缓、A型地中海贫血症、生殖器异常和面部畸形。CREB结合蛋白（CREB-binding protein, CBP）共激活子也与染色质重塑有关，并且已经被发现在一种罕见的人类综合征中存在突变。可以区分染色质激活与失活的特征之一是组蛋白乙酰化状态。组蛋白是真核生物核小体的主要结构蛋白，在将DNA组装成染色质过程中发挥着关键作用。在转录激活区域，染色质压缩率低，组蛋白高度乙酰化。鲁宾斯坦 - 塔比综合征就是一种症状是面部畸形、拇指宽大、脚趾宽大和精神发育迟缓的疾病。携带组蛋白乙酰化活性的核蛋白CBP编码基因中的突变可以导致该病。

（三）转录激活子突变与发育

一个典型的人类疾病中转录因子突变的例子是Pit-1。这个转录因子的特征是有一个

POU型同源结构域。它在腺垂体中表达，并且是分泌生长激素（growth hormone，GH）、泌乳激素（prolactin，PRL）以及促甲状腺激素（thyroid stimulating hormone，TSH）的细胞分化和生存所必需的。Pit-1突变导致联合垂体激素缺乏（combined pituitary hormone deficiency，CPHD）的智力迟钝。已经有部分人被查出携带CPHD及Pit-1突变。

（四）转录激活子突变与癌症

*p53*是一种肿瘤抑制基因。在所有恶性肿瘤中，50%以上会出现该基因的突变。由这种基因编码的蛋白质（protein）是一种转录（transcription）因子，其控制着细胞周期的启动。许多有关细胞健康的信号向p53蛋白发送。关于是否开始细胞分裂就由这个细胞决定。如果这个细胞受损，又不能得到修复，则p53蛋白将参与启动过程，使这个细胞在细胞凋亡（apoptosis）中死去。有p53缺陷的细胞没有这种控制，甚至在不利条件下继续分裂。像所有其他肿瘤抑制因子一样，*p53*基因在正常情况下对细胞分裂起着减慢或监视的作用。细胞中抑制癌变的基因"*p53*"会判断DNA变异的程度，如果变异较小，这种基因就促使细胞自我修复，若DNA变异较大，"*p53*"就诱导细胞凋亡。

*p53*基因突变后，由于其空间构象发生改变，失去了对细胞生长、凋亡和DNA 修复的调控作用，*p53*基因由抑癌基因转变为癌基因。

*p53*基因与人类50%的肿瘤有关，目前发现的有肝癌、乳腺癌、膀胱癌、胃癌、结肠癌、前列腺癌、软组织肉瘤、卵巢癌、脑瘤、淋巴细胞肿瘤、食管癌、肺癌、成骨肉瘤等，人类肿瘤中*p53*突变主要在高度保守区内，以175、248、249、273、282位点突变最高，不同种类肿瘤不同，如结肠癌和乳腺癌有相似的流行病学（包括地区分布和危险因素），但*p53*突变谱并不一致。结肠癌G：C、A：T转换占79%，而且50%以上转换突变发生在第3~5结构域的CpG位点。在乳腺癌中，只发现13%的转换在CpG位点。此外，G-T颠换在乳腺癌占1/4，但在结肠癌十分罕见。淋巴瘤和白血病的*p53*突变方式与结肠癌相似，即大部分突变为CPG位点的 转换，G→T颠换较低，A：T→G：C在A：T位点突变较高。伯基特淋巴瘤与其他B细胞淋巴瘤和T淋巴细胞恶性病变的*p53*突变谱相似，但伯基特淋巴瘤的转换突变较高。在非小细胞肺癌中G：C→T：A最普遍，食管癌颠换率很高，与肺癌不同的是，G：C和A：T位点有相似的突变率。我国启东地区50%为249癌码子的G→C、G→T颠换，而南非肝癌80%为G→T 颠换。骨肉瘤中*p53*突变率为75%，主要集中在5~9外显子。

第六节

应 用 实 例

Section 6 Application

一、结合序列和表达谱数据识别组合式转录调控模式和调控网络

细胞生命活动复杂性的基础是基因的表达。由信号分子、转录因子以及转录因子的靶基因所构成的调控网络是组织细胞复杂性的系统形式。mRNA的表达水平由非编码区上的顺式转录元件的逻辑输入信号决定。这些顺式转录原件形成的逻辑信号最终影响到细胞生物学过程,包括生理适应性,细胞多样性的产生以及形态发育等。由全基因组全局方法与技术,包括计算方法、分析转录调控网络的结构与动态特性是该分析解决方案的主要内容。全基因组分析已经证明了重要的转录网络组织是由mRNA水平上具有共表达模式的基因构成,也就是说许多生物学过程是由基因产物的同时性参与完成的。比较基因组学分析也得到多物种间这些基因调控结构序列的保守性。基于模式识别的算法识别调控序列已经应用于单细胞酵母中。可以针对不同数据源和需求分析调控网络:

本例采用统计学模型系统解决基因组范围的基因表达谱下的复杂调控模式,包括识别决定基因表达调控的DNA序列上的调控元件及其空间方位(定位与方向位置);识别组合式调控模式在不同条件下的功能。利用贝叶斯统计模型基于序列特征,更可以推测出基因表达谱特征,并将其与真是数据做比较得到该方法的准确度和解释力。

分析步骤:

1. 根据表达谱数据得到差异的基因集合(Gene Set)。
2. 采用Force-directed placement算法计算高度相关性的共表达基因。
3. 衡量非编码序列特征影响基因表达的程度。
4. 在表达谱数据集中的不同条件下,全局地计算组合式转录调控元件的规律模式。
5. 计算获得预测的转录调控元件序列motif;定位与方位方向。
6. 获取条件特异、空间特异,或时间序列上时间特异点上的转录调控模式和机制。

二、基于启动子结构元件的组合模式识别调控网络

转录调控的组合式调控具有重要的生物学意义,例如细胞可以使用多个不同的转录因子的组合参与多种不同条件下生物反应。本例基于生物基因的启动子序列中的motif组合,并结合芯片表达谱数据预测新的motif和motif之间的关联,进而构建特定条件下的转录调控网

络。另外可以预测发现是否在调控网络中具有不同生物功能模块的相互交联(Cross–Talk)。

分析步骤:

1. 构建已知和预测的启动子基序–motif; 预测motif使用AlignACE算法。
2. 发现所有motif的组合情况以及对应的基因。
3. 对具有motif组合的基因集(Gene set)计算表达一致性得分(expression coherence score)。
4. 识别具有统计显著性的协同性motif组合。
5. 根据motif协同性构建motif map 以及调控网络。
6. 同时比较单独motif和组合motif在表达谱上的效果。

结果形式:

1. 针对不同motif组合,计算motif组合之间的相对距离与出现频度分布图(图7–17); 以及motif协同性的方向偏好性(orientation bias),通过真实情形与随机模拟做对比计算协同

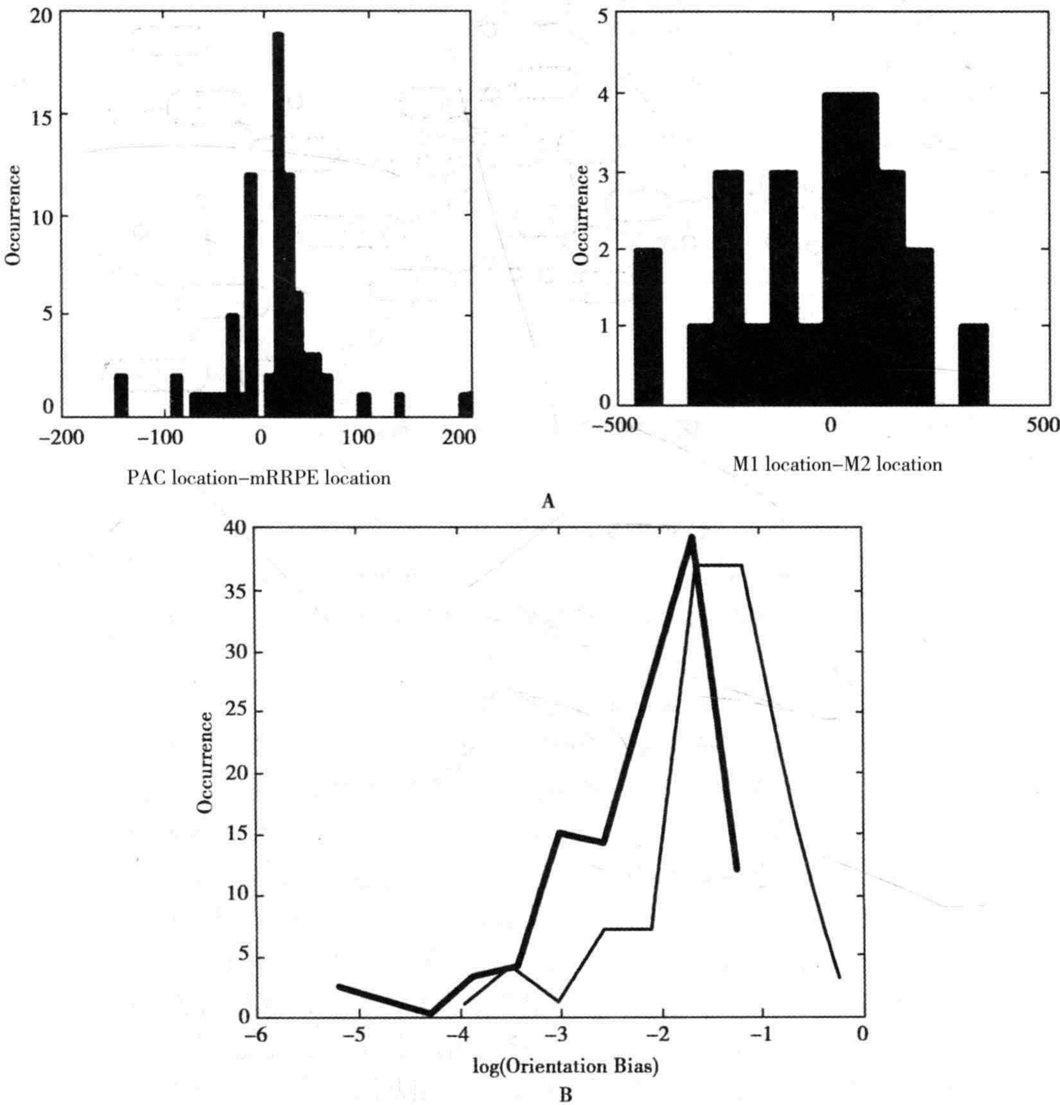


图7–17　motif组合之间的相对距离与出现频度分布图

motif的方向偏好性显著性，即Orientation bias score；显著性通过 Wilcoxon rank sum 统计学检验。另外，可视化协同性motif的共表达情况。

2. 构建全局motif协同性图谱（global motif synergy map）（图7-18）。已知和预测的motif之间的边表示组合协同性存在，边的*P*值计算说明协同性的可靠性。*P*值<<*P*0阈值的mitif组合才会最终可视化显示出来作为结果使用。

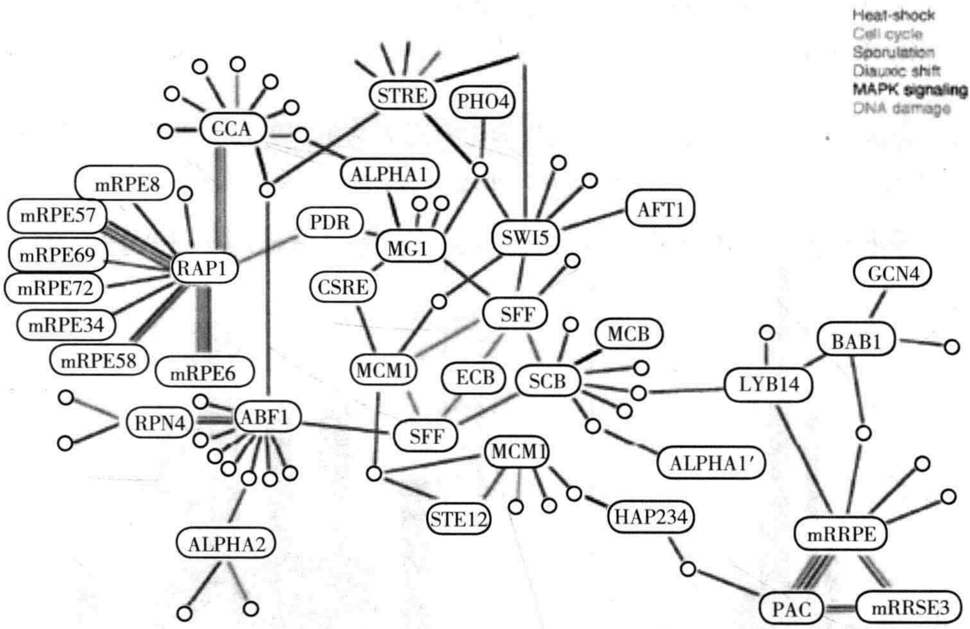


图7-18　全局motif协同性图谱

3. 探索计算motif与表达谱特征的关联（图7-19）。除了计算motif之间的关联组合，我们也要计算每个组合中的motif对表达谱的影响。我们进而可以识别是哪些motif组合对（motif pairs）或共享哪些motifs的基因对表达谱产生效应。协同的motif和在每个条件、样本下表达程度对应起来。

基于TF结合数据和表达谱数据识别遗传调控模块：

该例目的是识别基因共同被转录调控的模块（gene module）。基因模块被定义为具有共表达模式的以及共同被一组转录因子（TF set）转录的基因集（gene set）。整合转录因子TF和调控的基因模块将重构出转录调控网络（regulatory network）。该例并未假设在特定的基因模块中的基因表达模式直接受到调控模块转录因子的表达模式影响。因为在很多情况下，转录因子受到转录后修饰，而且该例并未能从表达谱中观测到基因产物即蛋白质的水平。因此，该例基于转录因子与DNA结合的组学数据，包括CHIP-chip，CHIP-seq等，以及基因表达谱数据（expression data），使得两个类型的组学数据相互补。该例首先基于TF-DNA结合强度的P值，选择被一组TF结合并且具有共表达模式的基因集。通过适当的放低结合强度的P值尽可能地将潜在共调控的靶基因考虑进来，建立一组基因共转录调控模块。

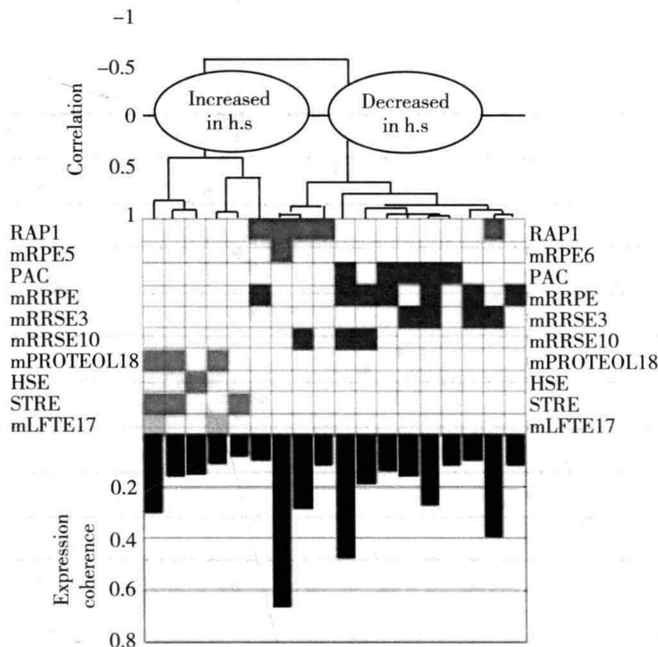


图7-19 **motif与表达谱特征的关联结果**

1. 共调控基因模块和重构的调控网络图(图7-20)。

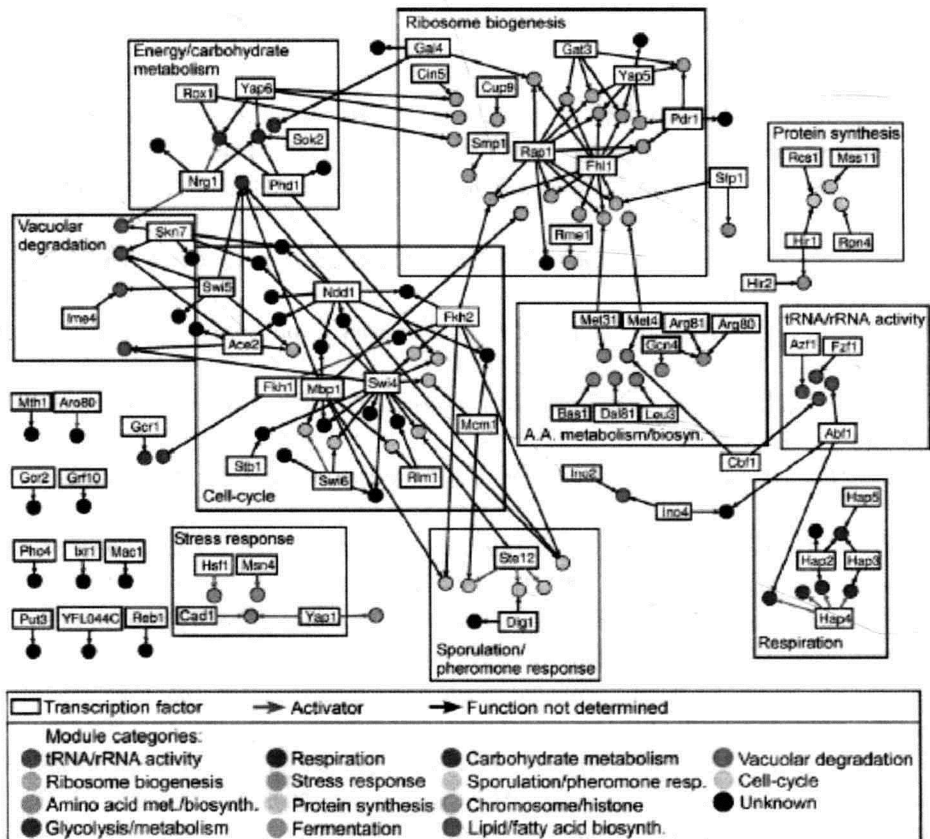


图7-20 **共调控基因模块和重构的调控网络图**

2. Motif 富集度分析(图7-21)。

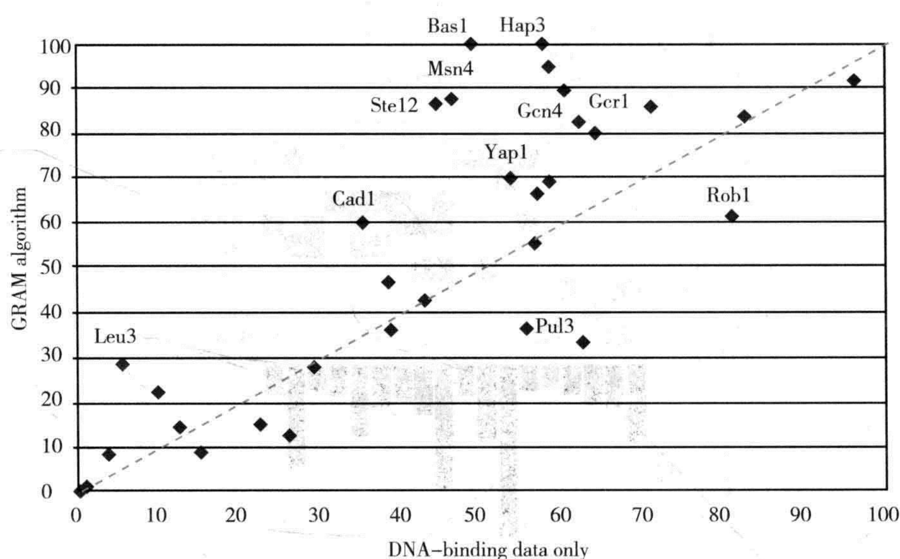


图7-21 **Motif富集分析**

（智 慧）

参考文献

1. Borenstein E, Kupiec M, Feldman MW, et al. , *Large-scale reconstruction and phylogenetic analysis of metabolic environments.* Proc Natl Acad Sci U S A, 2008. 105(38): 14482–14487.

2. Jerard Hurwitz. *The discovery of RNA polymerase.* Journal of Biological Chemistry, 2005. 280(52): 42477–42485.

3. Herr AJ, Jensen MB, Dalmay T, et al. , *RNA polymerase Ⅳ directs silencing of endogenous DNA.* Science, 2005. 308(5718): 118-120.

4. Wierzbicki AT, Ream TS, Haag JR, et al. , *RNA polymerase Ⅴ transcription guides ARGONAUTE4 to chromatin.* Nat. Genet. , 2009. 41(5): 630-634.

5. Bucher P, Trifonov EN. , *Compilation and analysis of eukaryotic POL Ⅱ promoter sequences.* Nucl. Acids Res. , 1986. 14 : 10009–10026.

6. Steen Knudsen. *Promoter 2.0: for the recognition of Pol Ⅱ promoter sequences.* Bioinformatics. , 1999. 15 : 356–361.

7. Halees AS, Leyfer D, Weng Z. , *Promoser: A Larger-scale Mammalian Promoter and Transcription Start Site Identification Service.* Nucl. Acids. Res. , 2003. 31 : 3554–3559.

8. Reese MG, *Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome.* Comput Chem . , 2001. 26(1): 51–6.

9. Bajic VB, Tan SL, Suzuki Y, , et al. , *Promoter prediction analysis on the whole human genome.* Nature Biotechnology. , 2004. 22 : 1467–1473.

10. Bajic VB, Tan SL, Suzuki Y, , et al. , *Computer model for recognition of functional transcription start sites in RNA polymerase Ⅱ promoters of vertebrates*, J. Mol. Graph. Model. , 2003. 21 : 323–332.

11. Bajic VB, Seah SH. , *Dragon Gene Start Finder identifies approximate locations of the 5' end of genes*, Nucleic Acids Res. , 2003. 31 : 3560–3563.

12. Bajic VB, Seah SH. , *Dragon Gene Start Finder: an advanced system for finding approximate locations of the start of gene transcriptional unites*. Genome Res. ,2003. 13 : 1923–1929.
13. Reese MG. . *Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome*, Comput. Chem. ,2001. 26 : 51–56.
14. Knudsen S. *Promoter 2. 0: for the recognition of Pol Ⅱ promoter sequences*, Bioinformatics,1999. 15 : 356–361.
15. Davuluri RV, Grosse I, Zhang MQ. , *Computational identification of promoters and first exons in the human genome*. Nat. Genet. ,2001. 29 : 412–417.
16. Ioshikhes IP, Zhang MQ. , *Large-scale human promoter mapping using CpG islands*. Nat. Genet. ,2000. 26 : 61–63.
17. Solovyev VV, Shahmuradov IA. , *PromH: Promoters identification using orthologous genomic sequences*. Nucleic Acids Res. ,2003. 31 : 3540–3545.
18. Down TA, Hubbard TJ. , *Computational detection and location of transcription start sites in genomic DNA*. Genome Res. ,2002. 12 : 458–461.
19. Ponger L, Mouchiroud D. , *CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences*. Bioinformatics,2000. 5 : 380–391.
20. Scherf M, Klingenhoff A, Werner T. , *Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach*. , J. Mol. Biol. ,2000. 297 : 599–606.
21. Ohler U, Stemmer G, Harbeck S, et al. , *Stochastic segment models of eukaryotic promoter regions*, Proc. Pac. Symp. Biocomput. ,2000. 5 : 380–391.
22. Solovyev V, Salamov A. , *The Gene-Finder computer tools for analysis of human and model organism genome sequences*, In Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology, 1997 : 294-302.
23. Zhang M. Q. , *Identification of human gene core promoters in silico*, Genome Res. ,1998. 8(3): 319–326.
24. Audic S, Claverie JM. , *Detection of eukaryotic promoters using Markov transition matrices*. Comput. Chem. , 1997. 21(4): 223–227.
25. Bajic VB, Seah SH, Chong A, , et al. , *Dragon Promoter Finder: recognition of vertebrate RNA polymerase Ⅱ promoters*, Bioinformatics,2002. 18(1): 198–199.
26. M. G. Reese, F. H. Eeckman, , *Time-delayed neural network for eukaryotic promoter prediction*, unpublished, 1999.
27. Hutchinson G. B. *The prediction of vertebrate promoter regions using differential hexamer frequency analysis*, Comp. Appl. Biosci. ,1996. 12 : 391–398.
28. Eponine, http: //www. sanger. ac. uk/Users/td2/eponine/, http: //servlet. sanger. ac. uk: 8080/eponine/
29. Ramana Davuluri, Ivo Grosse , Michael Zhang, http: //rulai. cshl. edu/tools/FirstEF/
30. Pedersen AG, Baldi P, Chauvin Y, et al. , *The biology of eukaryotic promoter prediction - a review*, Computers & Chemistry,1999,23 : 191–207.
31. Y. V. Kondrakhin, A. E. Kel1, N. A. Kolchanov, , et al. , *Eukaryotic promoter recognition by binding sites for transcription factors*, Comput. Appl. Biosci. ,1995,11 : 477-488.
32. Frech K, Danescu–Mayer J, Werner T. , *A novel method to develop highly specific models for regulatory units detects a new ltr in genbank which contains a functional promoter*, J. Mol. Biol. ,1997,270 : 674-687.
33. Chen QK, Hertz GZ, Stormo GD, *Promfd 1. 0: a computer program that predicts eukaryotic pol Ⅱ promoters using strings and IMD matrices*, CABIOS,1997,13 : 29-35.
34. Prestridge DS. *Prediction of pol Ⅱ promoter sequences using transcription factor binding sites*, J. Mol. Biol. , 1995,249 : 923-932.

第八章
CHAPTER 8

生物分子网络与通路

BIOLOGY MOLECULAR NETWORK AND PATHWAY

在过去的一个世纪中,分子生物学的很多领域只是研究少数几个基因、蛋白或分子的功能,但是,现在人们已经越来越意识到很多生物功能不只由几个分子或基因所控制的。相反,几乎所有的生物特征都是由细胞内众多成分(如DNA、RNA、蛋白质、酶和代谢子等)之间复杂的相互作用所引起的。因此,人们必须要在成千上万个生物分子组成的复杂系统的层面上予以认识,而不仅仅研究少数几个基因的功能。揭示数量巨大的生物大分子及其间的相互作用如何在复杂的生存环境中行使生物学功能,需要研究者采用不同于传统生物学研究手段的新技术。这样复杂的系统可以自然地模拟作人们很熟悉的概念:网络。生物分子网络与其他网络在现实世界中普遍存在。例如,近年来我国修建的高速公路将众多城市连接为一个巨大的网络,城市作为网络中的节点通过公路与其他城市连接在一起。互联网本身也是一个巨大的网络,网络服务器、个人计算机和其他计算设备被通讯线路连接在一起,通过网络中节点之间的连接,实现了全球计算机间的高速通讯与信息资源共享。在我们的周围还有人或者群体作为节点被多种多样的关系关联起来的社会学网络。在生物医学领域,各种复杂疾病,如各种癌症、糖尿病、高血压、精神分裂症等的发生和发展同样大多由于细胞内部多个分子、基因、蛋白的改变而影响正常的生物学过程。细胞内部的各基因、蛋白间,彼此相互作用进而形成复杂的蛋白质网络、基因表达网络、信号传导网络、转录调控网络、代谢网络等。因此,基于生物学网络的疾病相关研究中,研究者们通常利用网络分析技术,从系统角度揭示复杂疾病的产生和发展规律。本章将介绍生物学网络分析在系统生物学中的应用。通路作为生物学网络的一种重要类型,本章将对基于网络的通路分析进行详细介绍和应用实例展示。

第一节

网络的基本概念

Section 1 Description of Network

现实世界的复杂系统，尤其生物系统中，包含很多不同层面和不同组织形式的网络。生物系统中的网络通常由许多参与不同生物过程的分子元件组成，其中最重要的元件是基因、代谢子和蛋白质。但对"系统"而言，关键不是这些元件本身，而是元件之间的关联关系。基因、代谢子和蛋白之间并不是彼此孤立的，细胞内部的各基因、代谢子和蛋白间，彼此相互作用进而形成复杂的蛋白质网络、基因表达网络、信号传导网络、转录调控网络、代谢网络等。此外，生物学网络的发展使得各种典型分析策略和分析方法迅速扩展到其他由生物系统数据推导出的衍生网络中，如疾病基因网络。为了能够清晰地理解与分析网络，我们首先介绍网络的基本概念。

一、网络的定义

网络（network）通常可以用数学模型中的图表示，如G=（V, E）。其中 V 是网络的节点集合，E 是边集合。例如：每个蛋白质可以是图中的节点；那么蛋白质相互作用关系可以构成边集合。如果V中的两个节点v1与v2之间存在一条属于E的边e1，则称边e1连接v1与v2，或者称为v1连接于v2，也可以称作v2是v1的邻居。

二、有向与无向网络

根据网络中的边是否具有方向性或者说连接一条边的两个节点是否存在顺序，网络可以分为有向网络与无向网络，边不存在方向性为无向网络（undirected network），如图8-1A所示。否则为有向网络（directed network），如图8-1B所示。生物分子网络的方向性取决于其所代表的关系，如调控关系中转录因子与被调控基因之间存在顺序关系的，因此转录调控网络是有向网络，而基因表达相关网络中的边代表的是两个基因在多个实验条件下表达的高相关性，因此是无向的。

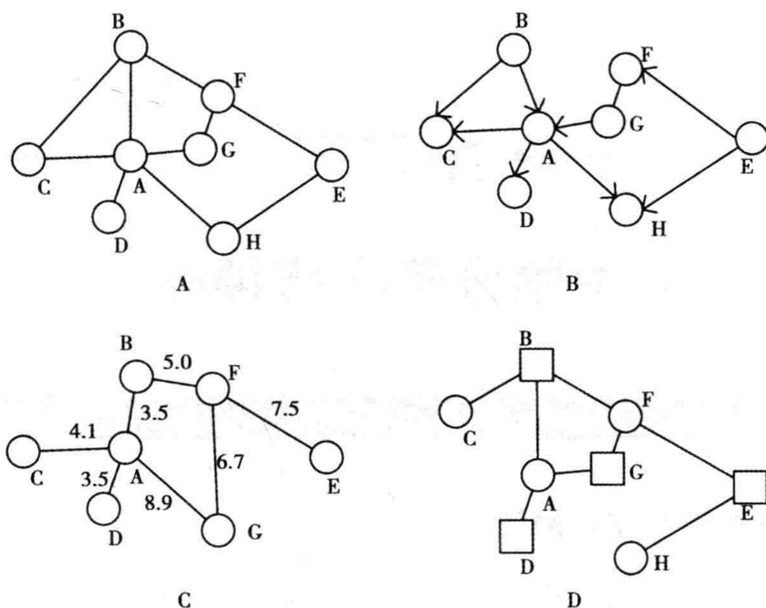


图8-1 网络分类图

A.无向网络；B.有向网络；C.加权网络；D.二分网络

三、加权与无权网络

如果网络中的边都被赋予相应的数值，这个网络就称为加权网络(weighted network)，所赋予的数值称为边的权重，如图8-1C所示。权重可以用来描述节点间的距离、相关程度、稳定程度等各种信息，含义依赖于网络和边本身所代表的意义。网络中的边权重是网络中普遍存在的一种现象。如交通网中，连接两个城市(节点)的道路(边)一般具有不同的长度，而在蛋白质相互作用网络中，蛋白质之间相互作用有强弱之分。网络中边权重的引入可以定量的分析网络系统，使结果精度得以显著提升。但分析的难度和计算量也将成倍加大。因此，最常用的网络仍然是无权网络。如果网络中各边之间没有区别，可以认为各边的权重相等，称为无权网络(unweighted network)。

四、二分网络

当网络中的节点能够分为两个互不相交的集合，并且所有的边都由不同集合的节点之间连接构成时，称这样的网络为二分网络(bipartite network)，如图8-1D所示。生物学网络中二分网络的现象非常普遍。例如，药物分子与其靶蛋白的结合关系、疾病与疾病基因的关系、疾病与通路关系、转录因子与靶基因绑定关系、microRNA与靶基因关系等都可以用二分网络模型表示和分析。

第二节

生物分子网络种类

Section 2　Classification of Biology Molecular Network

一、蛋白质相互作用网络

蛋白质相互作用网络是以蛋白质作为节点，参与同一代谢途径、生物学过程、结构复合体、功能关联或蛋白质间的物理接触作为边的网络，如图8-2A所示。目前来讲，蛋白质互作网络是被研究最充分的生物分子网络之一。蛋白质是组成生物体并行使生物功能的重要生物大分子。蛋白质通过相互作用构成网络来参与生物信号传递、基因表达调节、能量和物质代谢及细胞周期调控等生命过程的各个环节。因此，蛋白质互作网络对于理解细胞网络结构及功能，以及疾病发生发展的基础至关重要。

研究人员主要从生物实验检测和计算机预测两个角度来研究蛋白质相互作用。实验检测技术主要有免疫共沉淀（co-immunoprecipitation）、酵母双杂交（yeast two Hybrid，Y2H）和串联亲和纯化–质谱（tandem affinity purification – mass spectrometry，TAP-MS）技术。免疫共沉淀技术主要是在自然状态下，利用抗体抗原反应（western印迹法）检测与目标蛋白互作的其他蛋白，由此确定互作关系。它是当前最为可靠的蛋白互作检测技术。但无法检测短暂时间不稳定的蛋白互作关系，另外也需要预先确定待检测的互作关系用于准备相应的抗体，因此检测的效率比较低，从而无法应用于大规模的互作检测。酵母双杂交技术是根据酵母的某些转录因子（如GAL4）拥有DNA结合域和转录激活结构域，并且两个结构域空间接近时表现转录活性的特点，检测蛋白质互作的技术。通过载体转染、表达融合蛋白、报告基因表达等步骤，判断待检测蛋白质之间是否存在互作关系。酵母双杂交技术不仅用来研究哺乳动物蛋白质之间的互作关系，还可以用来研究高等植物蛋白质之间的互作。该技术的优点是检测通量高，但缺点是检测结果的假阳性互作较高。串联亲和纯化–质谱技术首先通过免疫共沉淀反应或串联亲和纯化反应得到含有目的蛋白的蛋白质复合物，然后用质谱分析或蛋白测序来鉴定复合体的各个组分。该检测技术的可靠性高于酵母双杂交技术，同时检测通量也高。但是仍不适用于检测瞬时的蛋白质互作。

随着生物信息学方法的不断发展，人们开发了许多的蛋白质互作预测技术。例如，利用蛋白质同源性预测、蛋白质结构模式等方面信息预测蛋白质互作。此外，也出现了Bayes网络等机器学习技术整合多种数据源的信息，预测蛋白质互作的网络技术。从生物实验检测和计算机预测获得的蛋白质相互作用信息的快速增长，产生了大量的蛋白质互作数据库。

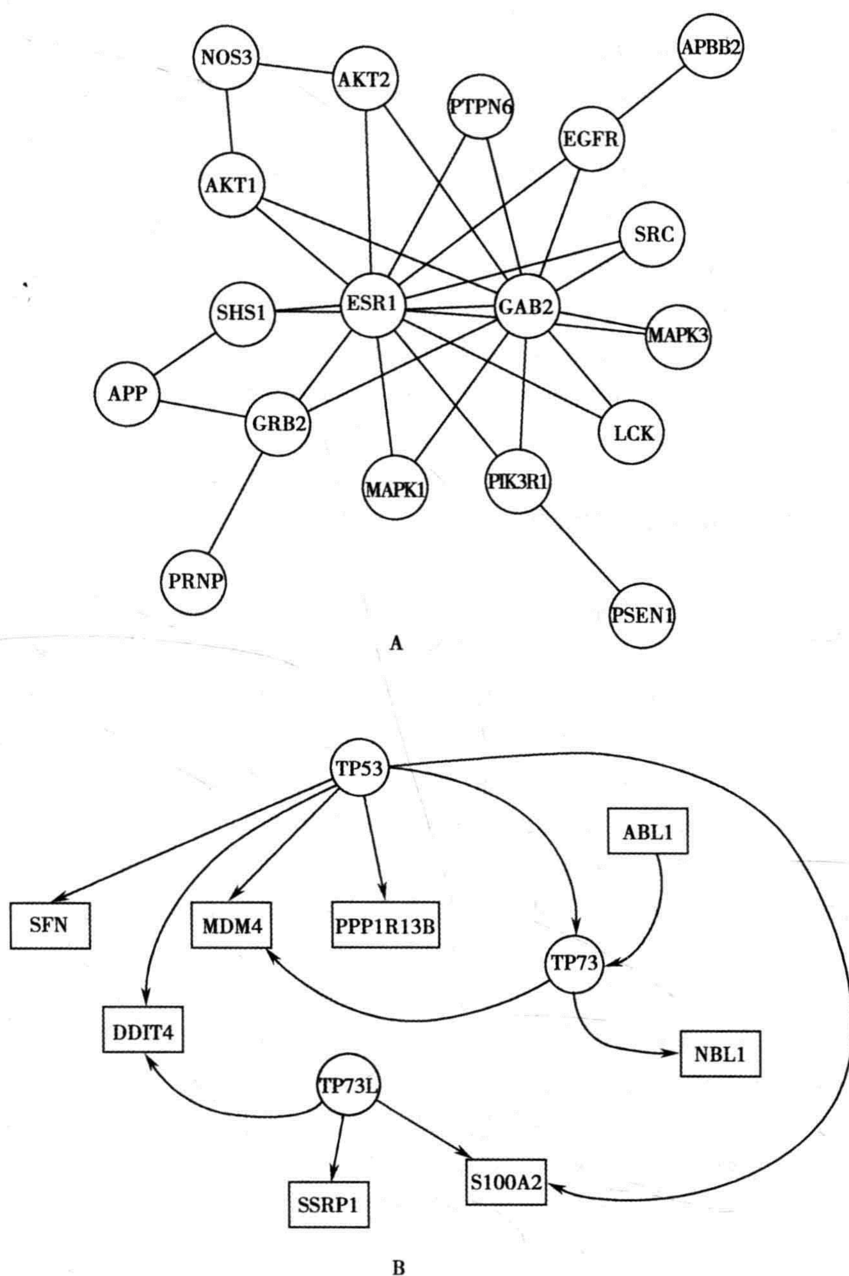


图8-2 A蛋白质相互作用网络；B 基因调控网络

这些数据库构成了最为庞大的生物学网络资源。这些资源对于理解生物系统中蛋白质的工作原理，了解疾病等特殊生理状态下生物信号反应机制、蛋白间的功能联系都有重大意义。下面列出了一些常被人们使用的数据库，包括：

1. STRING(http: //string-db.org/)数据库　该数据库中不仅存储了已知实验证实的蛋白质互作数据，还存储了预测的蛋白质互作数据。这些互作包括直接物理互作和间接互作(如：功能互作)。这些信息主要来源于四个主要途径，包括：基因组信息、高通量实验、基因共表达信息和先验知识。目前该数据库存储了来自于1133个物种的互作信息。用户可以通过网

络查询工具对蛋白质互作信息进行查询，也可以对互作信息进行下载。

2. BIND（http: //bind.ca/）数据库　该数据库主要记录蛋白质互作在内的生物分子间的相互作用信息。该数据库内收录的信息分为经过人工检测评价的高可信互作信息和高通量技术得到的互作信息。

3. HPRD（http: //www.hprd.org/）数据库　该数据库仅收录人类数据，是一个包含了蛋白注释信息，蛋白转录后修饰以及蛋白互作等多种信息的人类综合性数据库。HPRD数据库已经成为文献挖掘方法获取蛋白互作及转录后修饰的最大的数据库之一。

4. DIP（http: //dip.doe-mbi.ucla.edu/）数据库　该数据库包含人工检查评价的可靠互作信息和计算预测所获取的高通量数据。该数据库中支持多种物种。

5. MIPS（http: //www.helmholtz-muenchen.de/en/mips/）数据库　该数据库包含了多种蛋白质互作信息、其中蛋白质复合物信息较为全面。该数据库也支持多种物种。

二、基因转录调控网络

基因转录调控网络是以转录因子和受到它们调控的基因作为节点，以这些节点间的调控关系作为边的有向网络，如图8-2B所示。通过获得大量的基因转录调控数据可以直接构建复杂的基因转录调控网络。网络中的边可以依据转录因子是促进还是抑制受调控基因的表达，分为正调控和负调控两种边的关系类型。

目前检测基因调控的技术已比较成熟，主要包括染色质免疫沉淀技术（ChIP）和在此基础上发展起来的ChIP-chip芯片及ChIP-chip等技术。ChIP可以检测体内转录因子与DNA的动态作用。与体内足迹法、DNA芯片和分子克隆等技术相结合，ChIP技术已成为研究DNA与蛋白质相互作用的重要方法。ChIP-chip芯片是将生理状态下细胞内的蛋白和DNA结合在一起，利用超声波将其打碎，然后特异性地富集目的蛋白结合的DNA片段和纯化检测这些片段，最终获得蛋白与DNA作用的信息。另外，微阵列技术通过关联基因之间的表达水平也可推断基因调控关系。随着这些技术的发展，在短时间内可获得生物体基因调控的海量数据，这为研究和揭示基因及其产物之间的相互关系，特别是基因转录的调控机制奠定了基础。目前，许多数据库收集了大量的基因转录调控信息，一些流行的数据库包括：

1. TRANSFAC（http: //www.gene-regulation.com/pub/databases.html）数据库　该数据库提供转录因子以及它们在基因组上的结合位点信息。该数据库由SITE、GENE、FACTOR、CLASS、METHOD、REFERENCE等部分组成。此外，TRANSFAC数据库还与几个扩展库如PATHODB、S/MARTDB、TRANSPATH、CYTOMER库密切关联。

2. JASPAR（http: //jaspar.genereg.net/）数据库　该数据库是存储了真核生物中转录因子和DNA结合位点的最全面的数据库之一。JASPAR包括核心数据库和其他几个子数据库，是一个非冗余的数据库，其中包含的数据都经过严格的筛选，具有实验条件。

3. TRRD（http: //wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/）数据库　该数据库提供真核生物基因调控区结构、功能特性信息。转录因子结合位点、启动子、增强子、静默子以及基因表达调控模式等。

4. COMPEL（http: //compel.bionet.nsc.ru/）数据库　该数据库存储了许多复合转录元件，包括不同转录因子在位置关系上紧密相连的结合位点、结合部位之间的距离和先后顺序，以

及转录因子的三维空间结构。

5. TRED(http: //rulai.cshl.edu/cgi-bin/TRED/tred.cgi ? process=home) 数据库　该数据库是一个转录调控元件数据库，收集了实验证实的包含人类、小鼠、大鼠物种等哺乳动物顺式调控元件和反式作用因子、转录因子相关数据。该数据库不仅提供转录因子结合位点序列信息，还提供转录因子结合位点的基因组定位信息。数据经过人工校正，实验证实，具有一定可靠性，这些数据完全公开。

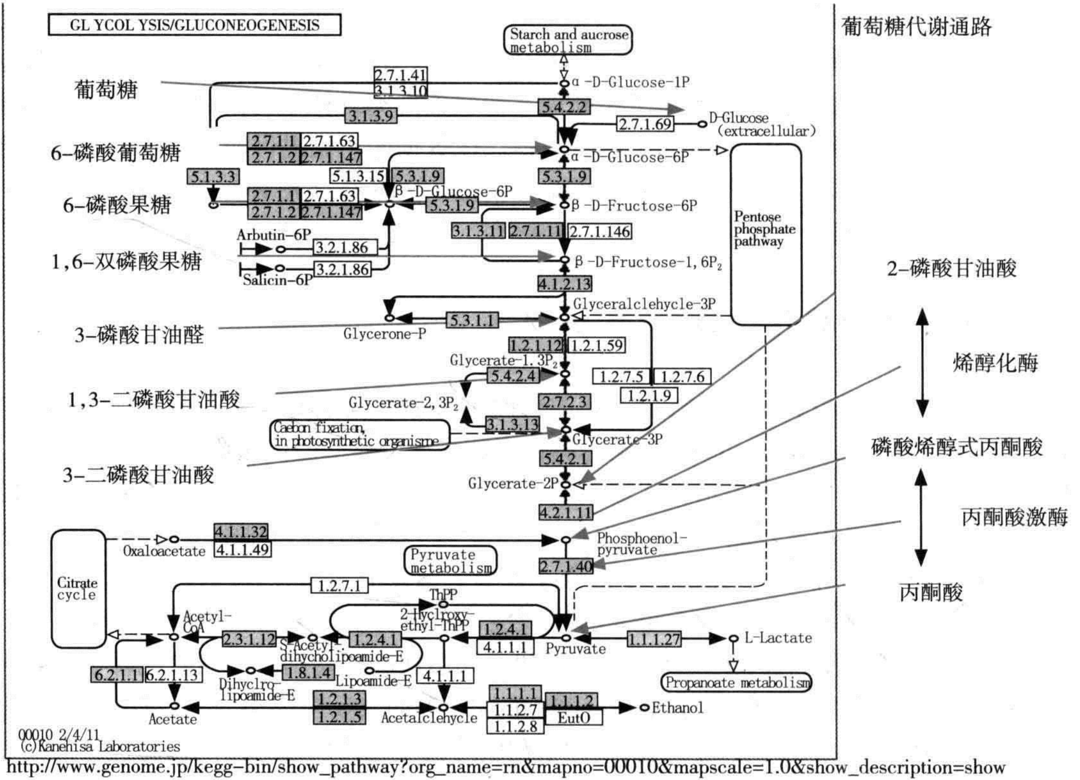
三、代谢和信号传导网络

细胞内代谢物在酶的作用下转化为新的代谢物过程中发生的一系列的生物化学反应形成了代谢通路(metabolic pathway)。葡萄糖代谢就是一个典型的代谢通路，如图8-3所示。这样的代谢通路可以自然地表示作代谢网络。与其他的代谢通路的联合又会形成更大的代谢网络。代谢网络包含代谢子、酶等生物分子之间的多种生理和化学反应，酶和代谢子在网络中可能多次出现，一个节点也可能对应多个生物分子，因此代谢网络与蛋白质互作网络等其他生物分子网络相比具有更大的复杂性。网络属于复杂的超图模型范畴。人们往往为了简化网络的复杂性，根据研究目的构建不同层次的代谢网络。当研究者不关心代谢反应中的酶和其他一些如提供能量与磷酸键的ATP等的共反应因子，就可以将网络转化为只包含主要代谢底物指向主要产物的代谢子网络。甚至忽略反应方向的情况也经常被许多研究者采用。基因组学和蛋白质组学的发展更使得研究者经常将代谢网络简化为强调基因和酶的网络，而弱化代谢子。一种常用的方法是转化代谢通路为以酶为节点，酶和酶之间如果通过生化反应直接共享至少一个代谢子，那么它们之间连接一条边。进一步，通过获得基因编码酶的信息，可以将网络转化为基因网络。除了这些，还有许多简化方法被研究者广泛使用。目前，一些软件也可处理代谢通路数据网络简化，如基于R语言的两个软件包iSubpathwayMiner(http: //cran.r-project.org/package= iSubpathwayMiner) 和KEGGgraph(http: //bioconductor.org/packages/2.4/bioc/html/KEGGgraph.html) 提供了多种方便的通路简化方法。

细胞通过将生物信号或刺激转换为其他生物信号最终激活细胞反应的过程形成了信号传导(signal transduction)，信号传导的过程中多个生物分子在酶作用下按照一定顺序发生一系列生理化学反应，由此形成信号传导通路(signal transduction pathway)。与代谢通路相似，信号传导通路可以自然的表示作信号传导网络。网络同样属于复杂的超图，并且网络中边的种类非常多。如图8-4所示，JAK-STAT信号通路中包含了激活、磷酸化、泛素化等多种信号作用信息。

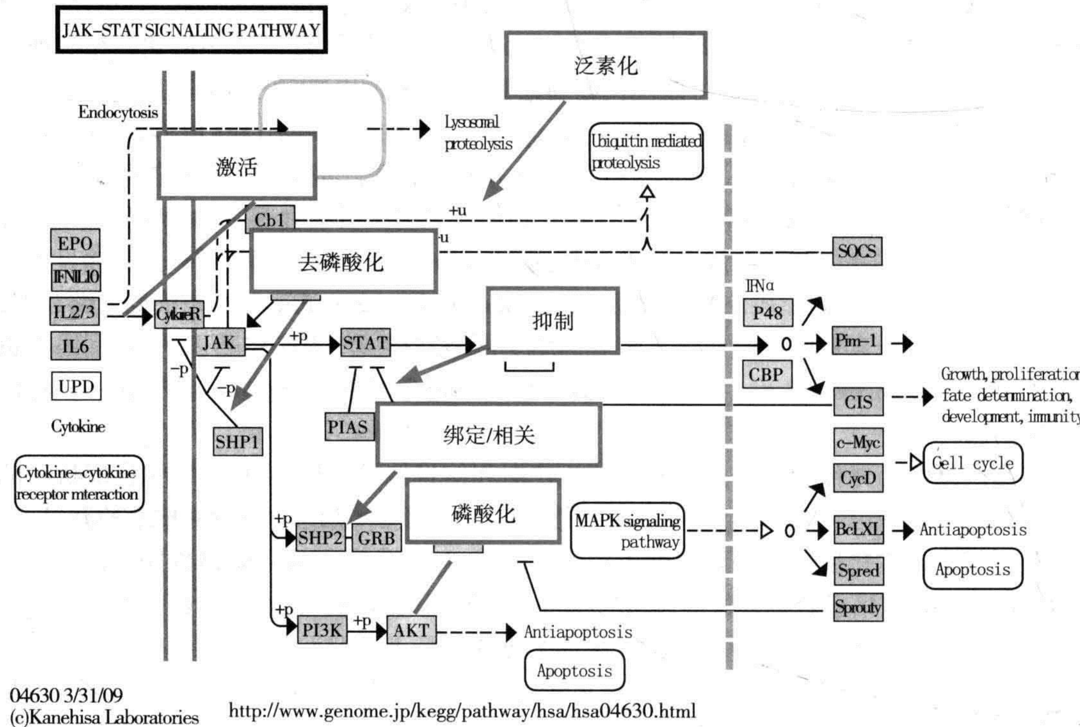
代谢网络和信号传导网络是研究和分析代谢和信号传导过程，疾病的发生发展机制的重要工具。随着基因组学、转录组学、蛋白质组学和代谢组学新的生物检测技术的开发，人们对生物细胞内生化反应的理解程度正在不断加深，各种代谢和信号通路数据也正以极快的速度增加。这些信息是构建代谢网络与信号传导网络的基础。目前这些信息被收集和整理到一些重要的通路数据库当中，主要的通路数据库包括:

1. KEGG数据库(http: //www.genome.jp/kegg/)　该数据库是关于基因、蛋白、酶、代谢子、药物、生化反应以及通路的综合生物数据库。该数据库实际由多个子数据库构成。最著名的当属通路(KEGG PATHWAY) 子数据库。它是目前最被广泛使用的通路数据库。



http://www.genome.jp/kegg-bin/show_pathway?org_name=rn&mapno=00010&mapscale=1.0&show_description=show

图8-3　KEGG数据库中的葡萄糖代谢通路



http://www.genome.jp/kegg/pathway/hsa/hsa04630.html

图8-4　KEGG数据库中的JAK-STAT信号通路

其中包含有上千个物种的代谢与信号传导通路信息。这些信息从生物学实验和文献中提取,并经过人工校正。实时更新的管理模式也使得人们能够从该数据库获得最新的通路数据。

2. Reactome数据库(http://www.reactome.org/)　该数据库是一个含多物种信息的通路数据库,存储了大量的代谢通路信息及生化反应信息,这些信息从生物学实验和文献中提取,并经过人工校正。该数据库中所有的生物过程中的反应以分层次的方式组织起来的,较低的层次对应着反应,较高的层次代表着通路。

3. WikiPathways数据库(http://wikipathways.org/)　该数据库是一个开放的共同协作的通路数据库平台。该数据库平台允许任何人创建新的通路数据,并由专业的生物学家进行校正,因此该数据库对现有通路数据库如KEGG, Reactome等进行了补充。虽然目前还不够强大,但该数据库的共同协作模式将极大地改善通路数据的规模和质量。

4. Pathway commons数据库(http://www.pathwaycommons.org/)　该数据库是一个包含了生物通路信息及蛋白互作信息的多物种综合数据库。它包含了来自Reactome、HumanCyc、HPRD等多个数据库的信息,因此可以作为获得公共通路数据库通路信息的一个接口使用。

5. PID数据库(http://pid.nci.nih.gov/)　该数据库是人类细胞信号通路数据库,存储了大量的信号通路和关键的反应以及各种分子互作。PID中包含了三个不同来源的数据,第一个来源是由NCI组织校正的通路,这种通路是从同行评议的文献中获得的;第二个来源来自Reactome数据库,第三个来源由KEGG数据库提供。

四、衍生网络

除了上面介绍的几种常见生物分子网络外,还有一些在它们的基础上衍生出来的生物网络。例如疾病基因网络、疾病通路网络、药物通路网络等。这些衍生网络都是利用基础网络信息和基本数据库资源构建出的新型网络,适用于分析各种具体的生物学问题,下面我们简要介绍几种衍生网络。

(一)疾病基因网络

遗传异质性和基因多效性是很常见的生物学现象。疾病的发生发展通常是由多基因突变造成的。随着实验技术的发展,人们对疾病的认识以及各类疾病与致病基因之间的关系的广泛理解,产生了复杂的人类疾病与基因之间的关系。疾病基因网络可以从全局的角度来分析人类疾病和致病基因之间的复杂关系。

为了构建疾病基因网络,我们可以从疾病相关数据库,例如OMIM(人类孟德尔遗传在线)数据库中获得人类疾病和疾病基因的相关信息,经过一定的筛选处理后,得到了疾病和基因数据,然后将被证实的疾病和基因作为节点,疾病和基因关系作为边,这样就构成了人类疾病基因网络.如图8-5A所示,网络中圆形代表人类相关疾病,方形代表相应的致病基因。人类疾病基因网络跟其他的基本网络一样,可以分析网络的基本性质,对该网络进行功能聚类、模块化等基本的网络分析。可以研究疾病基因在疾病发生发展过程中所起的作用以及从全局的角度分析某一种具体疾病或致病基因在网络中的交互作用,可以使我们加深对疾病基因在致病过程中生物学作用的理解以及从更全面的角度解释了疾病的发生和发展过程。

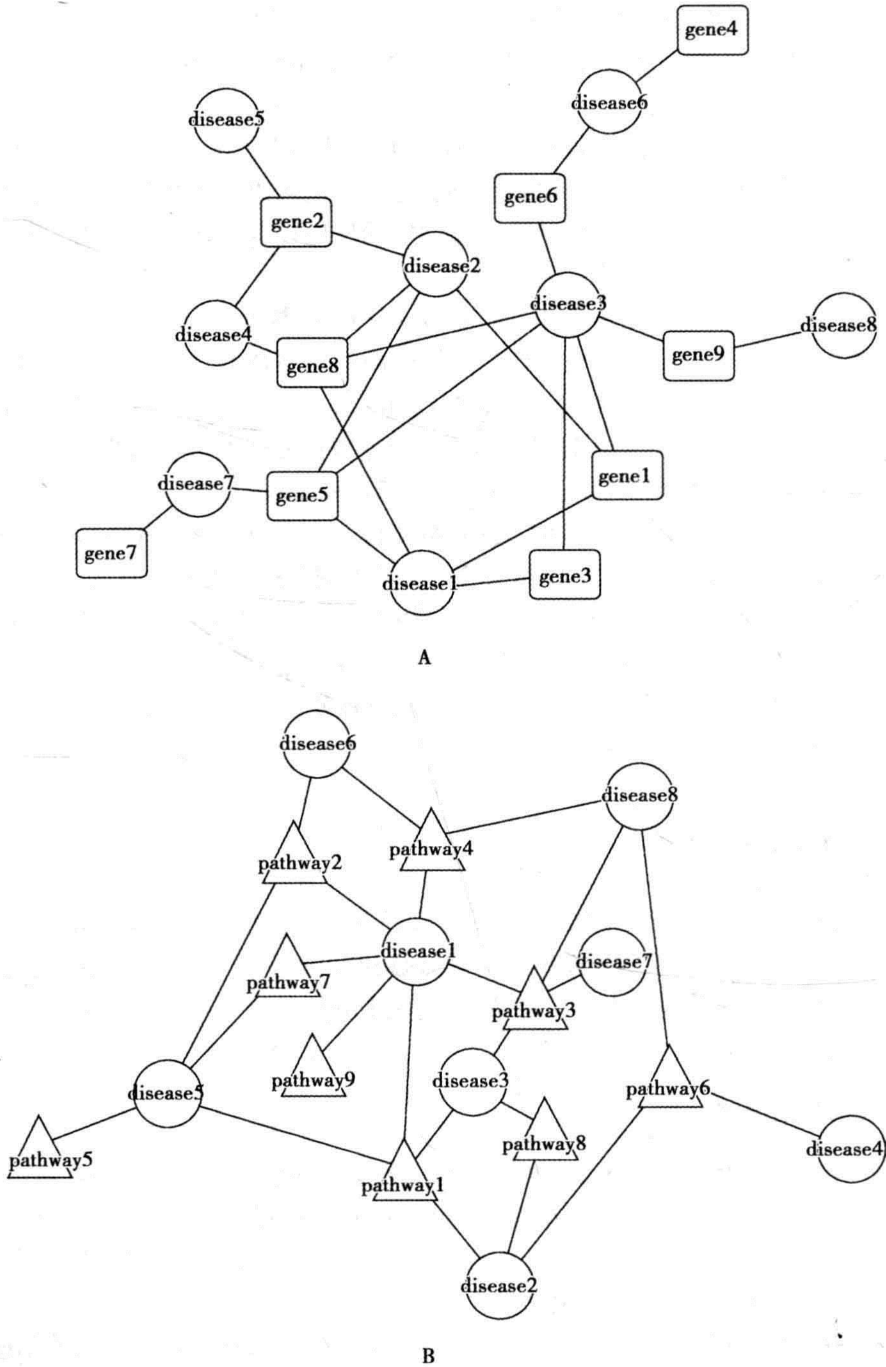


图8-5 A 疾病基因网络示意图；B 疾病通路网络示意图

(二) 疾病通路网络

基因或蛋白质并不是孤立的，他们通过互作形成代谢、调控等网络来行使生物学功能的，各种各样的通路网就是这些网络的典型代表，同样通路在发育、生长、衰老和死亡等一系列生物学过程中起了关键的作用。随着通路信息的逐渐完善，人们逐渐用各种方法从通路的角度来分析疾病，并发现通路在疾病的起始、进展和转化等过程中起了至关重要的作用。

当然，构建出一个全局的疾病通路网络是从整体上分析疾病与相关通路复杂关系的重要方法。

我们可以从疾病数据库（如OMIM）和通路数据库（如KEGG）分别获得疾病基因信息和通路基因信息，然后将每个疾病的所有疾病基因分别做基因富集分析。显著的富集通路（如：$p<0.01$）与该疾病就可以建立关联关系，最终构成疾病通路网络，图8-5B为疾病通路网络示意图，在网络中共有两类节点，三角形代表通路信息，圆形代表疾病，边相连代表着该疾病的发生发展和这条通路相关，这样的网络能够显示了人类各种疾病和疾病相关通路之间的复杂关系，通过对该疾病通路网进行网络分析，我们能全局性的了解各种通路以及在疾病的发生发展中一些规律，对人类疾病的机制解释和治疗有着指导意义。我们将在本章第五节中提供一个实例，进一步利用网络分析技术更精细的构建和分析疾病-通路网络。

（三）药物通路网络

药物的多靶点、多效应、多途径等特性给药物的开发和研制带来了很大的困难，人们也渐渐地发现大多数药物并不是作用于单个蛋白或基因产物而发挥功能的，药物发挥药效作用的过程相当复杂，近年来，人们把药物研究的视野逐渐转向通路，力求从通路的角度探讨药物的相关问题，跟疾病通路网络一样，药物通路网络可以作为一个从通路的角度出发研究药物的作用机制、药效以及副作用等问题的重要方法。同样，我们可以从多种数据库（CMap数据库、DrugBank数据库和KEGG数据库等）中获取通路和药物相关信息，例如CMap数据库就提供了每种小分子药物影响下的基因表达谱，从这样的数据中可以得到许多药物信息和每种药物影响的基因信息，与疾病通路网络类似，如果一种药物所影响的基因能显著富集在一个通路上，那么就把这个药物和这个通路连接起来，对于所有的药物和通路来说，就构建出药物通路网络。我们可以从中探寻两个同类药物（或两种不同药物）与通路的连接情况，同样也可以研究多个通路被单个或药物联合调控的现象。总之，结合药物的多靶点、多功能等特性，药物通路网络可以从全局上以通路的角度分析药物作用过程中的特征与性质，为实验人员和药物开发者提供了很好的思路，为新药的开发和研制奠定了一定的基础。

第三节 生物分子网络分析方法

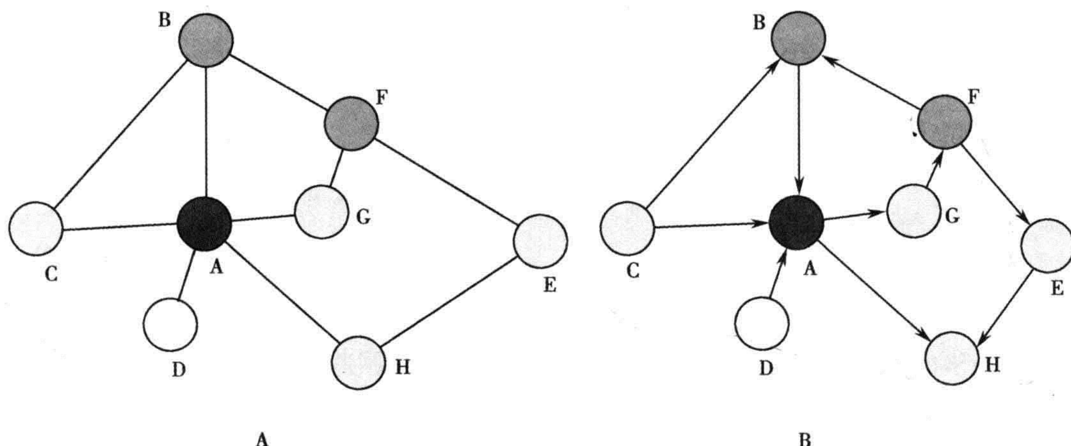
Section 3 Analysis Methods of Biology Molecule Network

一、网络的拓扑属性分析

网络的拓扑属性通过考察节点或边结构特征来描述网络全局及局部的特性，能够对网络进行初步探索。更重要的是，这些属性构成了深入分析网络的基本框架。通过进一步结合生物学信息和生物系统特点，网络的拓扑属性能够对深入理解生物系统及疾病的生物学机制起到关键的作用。下面介绍一些基本的网络拓扑属性分析测度。

（一）连通度

连通度或度（degree）是节点最基本的拓扑属性。某节点的度定义为网络中直接与该节点相连的所有边的数目，例如在图8-6A中节点A的连通度为5。如果网络是有方向的，我们通常还要定义两个不同的连通度的度量方式称为入度和出度。入度表示网络中直接指向该节点的所有边的数目，相反，出度表示网络中该节点直接指向其他节点的所有边的数目，例如在图8-6B中节点A的入度为3，出度为2。在本章中，我们用符号k表示连通度，k_{out}表示出度，k_{in}表示入度。连通度是非常重要的拓扑属性。尤其连通度高的节点，我们称之为中心节点（hub），即hub节点，与各类分子生物学功能、疾病发病机制等密切相关。癌症相关的基因也往往是hub节点。在蛋白质互作网络中，必需基因的翻译产物往往具有非常高的连通度。



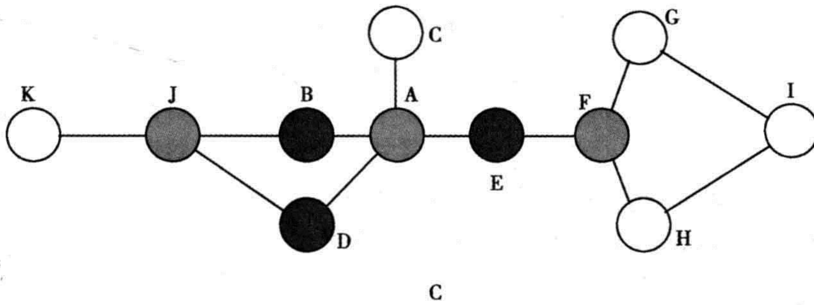


图8-6 网络拓扑属性的图例描述

(二) 聚类系数

对于无向网络来说，如果节点v_1连接节点v_2，节点v_2连接于节点v_3，那么节点v_3与节点v_1是否也会相连接？二者连接的可能性有多大？衡量网络中的这种现象的程度，可以使用聚类系数CC(clustering coefficient)。它能够测量部分节点间存在的密切连接程度。无向网络中，聚类系数定义如下：

$$CC_v = \frac{n}{C_k^2} = \frac{2n}{k(k-1)} \qquad (8\text{-}1)$$

公式中n代表节点v的所有k个直接邻居间存在的所有边的数目。因为n的最大数目可以由邻居节点的两两组合数$C_k^2 = \mathrm{k}(\mathrm{k}-1)/2$来确定，因此$CC$值的取值范围在[0，1]区间。如图8-6A所示，节点A有5个邻居{B，C，D，G，H}，邻居间仅仅有一条边连接，所以节点A的聚类系数$\mathrm{CC_A} = \frac{2\times1}{5\times(5-1)} = \frac{1}{10}$

对于有向网络，因为两个节点之间允许存在方向相反的边，此时聚类系数被标准化为：

$$CC_v = \frac{n}{P_k^2} = \frac{n}{k_{out}(k_{out}-1)} \qquad (8\text{-}2)$$

公式中k_{out}代表v的出度，n代表所有v所连接的节点相互之间存在的边数。在图8-6B中，节点A连接2个节点G和H，其间不存在连边，则节点A的聚类系数为$\mathrm{CC_A} = \frac{0}{2\times(2-1)} = 0$。而对于节点C出度也为2，连接节点B和A，其间存在一条边{B→A}，其聚类系数为$\mathrm{CC_C} = \frac{1}{2\times(2-1)} = \frac{1}{2}$

(三) 介数

节点的介数(betweenness)是该节点出现在其他节点间最短路径中的比例。介数越高，意味着在保持网络紧密连接性中节点越重要。节点v的介数B_v定义为：

$$B_v = \sum_{i\neq j\neq v\in V} \frac{\sigma_{ivj}}{\sigma_{ij}} \qquad (8\text{-}3)$$

其中，σ_{ij}指节点i到j的最短路径的数目，σ_{ivj}表示其中通过节点v的路径数目。两节点间的最短路径指连接它们的所有路径中最短的路径。例如在图8-6B中节点C到节点G的路径有l_1={C，B，A，G}和l_2={C，A，G}，但最短路径为l_2={C，A，G}。

（四）直径

网络的直径（diameter）是所有连通节点之间最短路径长度的最大值。网络的直径代表着整个网络紧密的程度，是衡量网络总体性质的指标。

（五）平均距离

网络的平均距离（average distance）是指网络中所有连通节点之间最短路径长度的平均值，同样用于代表整个网络的紧密程度。

（六）桥梁中心性

前面几个网络拓扑属性，包括度、介数和聚类系数都是反映了节点在整个网络中的中心地位，而桥梁中心性不仅能够反映节点在网络中的全局中心位置，同时还考虑了在网络中的局部特征。节点的桥梁中心性测度是介数中心性和桥梁系数的产物。对于网络中的节点v，其桥梁中心性定义为如下公式：

$$C_R(v)=BC(v)\times CB(v) \tag{8-4}$$

其中，BC（v）表示节点v的介数，如式8-3，CB（v）表示节点v的桥梁系数，某节点的桥梁系数决定了该节点处于连接度很高的节点间的程度，定义方式如下：

$$BC(v)=\frac{d(v)^{-1}}{\sum_{i\in}N(v)\frac{1}{d(i)}} \tag{8-5}$$

d（v）表示节点v的度，N（v）表示节点v的邻居节点集合，桥梁系数评估了邻居间的局部桥特征。

$C_R(v)$值越高预示着有越多的信息会经由节点v。如图8-6C所示，节点E、B、D的桥梁中心性都很高其中节点E最高$C_R(E)=0.45$。

将这些桥梁中心性值高的节点称为桥节点，这些桥节点位于网络当中聚集性相对较高的各个模块之间。总体来说，桥梁中心性不仅考虑了节点在网络中的全局中心位置，同时还考虑了在网络中的局部特征。

（七）易损性

网络中某个节点对于整个网络的连通性贡献有多大？如果将某节点删除，是否会对整个网络信息交流有影响？影响有多大？易损性能够衡量某节点对整个网络的信息传递影响程度。具体地，利用网络的全局特性来计算某节点对网络损坏的程度。网络全局特性是指网络节点间信息传递的效能，用下面的公式来计算：

$$E=\frac{1}{N(N-1)}\sum_{i\neq j}\frac{1}{d_{ij}} \tag{8-6}$$

其中，d_{ij}表示节点i与节点j之间的最短路径，N表示网络中节点的数目。

如果将节点v从网络中删除，那么整个网络的信息传输能力的破坏程度表示为：

$$V_i=(E-E_i)/E \tag{8-7}$$

其中，E表示原始网络的全局效能，E_i表示去除节点i后网络的全局效能。V_i越大，表示删除节点v后整个网络节点间的信息传输能力下降的越大，如果这类节点被删除了，那么整个网络中信息传输的效能必然会变差。如将图8-6C中的节点A删除后，整个网络将被分割为了两部分，信息交流效率将变差。

二、网络的无尺度特性分析

无尺度网络（scale-free network）指网络中度的分布符合幂率分布，即$p(k)\sim k^{-\gamma}$的网络。如图8-7A所示，为一个凹形的曲线。当将坐标转化为对数（log）坐标系后，分布接近为直线，如图8-7B所示。因此，无尺度网络中大部分节点的连通度较低，少数连通度非常高的节点使网络连接在一起。无尺度网络的网络直径相对较小，通常直径的大小正比于网络中节点数目的对数值的对数值，即log（log（N））。这样直径小的网络俗称小世界网络。

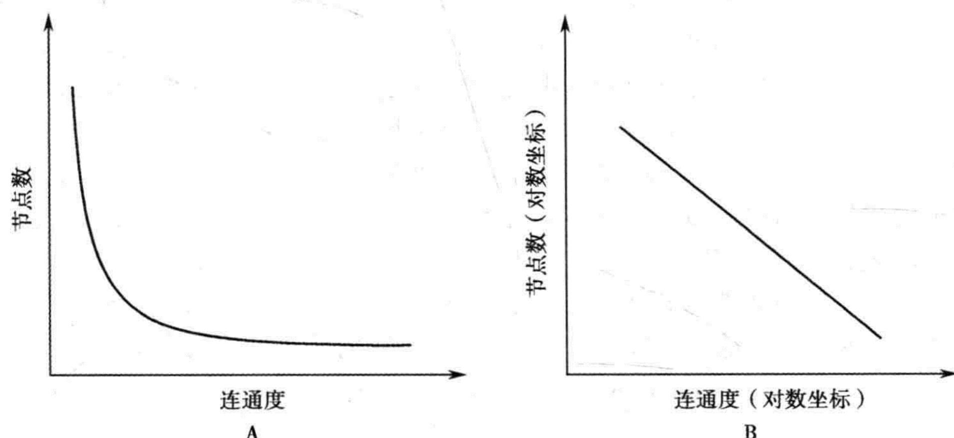


图8-7　无尺度网络度的分布

许多自然状态下的网络，如互联网和人际关系网络，都是无尺度网络。生物系统中，无尺度网络现象更加普遍。为了解释无尺度网络为何会成为生物分子网络的主要展现形式，Barabási和Albert提出了构建了形成无尺度网络的Barabási-Albert模型。根据这一模型，研究者模拟出蛋白质网络中出现无尺度特性的原因源于基因复制。高度连接的节点倾向于与发生复制的基因产物发生互作，从而获得额外的连接。这符合无尺度网络的两个特点：成长（growth）和优先连接（preferential attachment）。

成长性展现无尺度网络可以扩充规模。如网页的快速增加。优先连接表明网络的节点具有连接的优先级的区别，往往最初度大的节点可以在网络增长时形成更多的连接。如，著名的网站，更倾向于连接更多的网页和被其他新网站连接。因此，利用成长性和优先连接性质可知，度高的节点（hub）更倾向于是早期节点。研究表明，大肠埃希菌代谢网络中度高的几种分子，的确是最古老的代谢路径的一部分，而且它们的进化历

史悠久。基因复制很可能使得生物网络成长并进行优先连接,最终形成生物网络无尺度网络。

另外,无尺度网络具有强韧性,即对意外故障的抵抗能力强。例如,在计算机网络随机破坏个别的节点不会导致网络大面积的瘫痪。人体内基因也会随机的产生异常突变,但大多不会致死。正是因为无尺度网络大多数节点具有较小的度,随机损坏的个别节点往往是不重要的节点。因此,破坏网络的能力非常有限,体现在致病但不致死。对因特网和细胞而言,强韧性使得网络能够应付随机出现的异常。但是,生物学实验也显示,去除那些度高的蛋白质,经常导致细胞死亡。这说明,无尺度网络对hub节点的依赖,可能既有利也有弊。

三、网络的模序搜索

网络模序(motif)是指一类特殊的子网模式,即一组节点按照特定的顺序连接而成的结构,这类子网模式在网络中出现次数远超过随机情况。在生物学网络中,包含有大量的这些特殊的网络模序,搜索这些模序可以深入理解生物网络执行生物功能的基本形式,发现功能元件的功能关联关系。在有向网络中,人们发现了基因调控网络的前馈环(feed-forward loop),自调控环(auto-regulator loop)和单输入模序(single input motif)等一些非常重要的模序。在许多物种中,前馈环模序是一种非常常见的生物调控模式。如图8-8A所示,前馈环的一个例子:转录因子A调控转录因子B和基因C,而当转录因子B也调控基因C时,A、B和C形成前馈环结构。自调控环如图8-8B、C所示,由于调控机制可以为正向和负向,自调控环有2种不同的类型。单输入模序由同一个转录因子同时调控许多基因表达,而转录因子通常具有自调控性,所有调控方向都相同,且受调控基因不再受到其他元件的调控,如图8-8 D所示。该模序经常出现在大肠埃希菌(E.Coli)代谢通路相关的调控中。除上述模序外,研究者们还发现了许多其他的模序,如密集重叠调控(dense overlapping regulation)、多输入模序(multi input motif)和调控链(regulator chain)等。

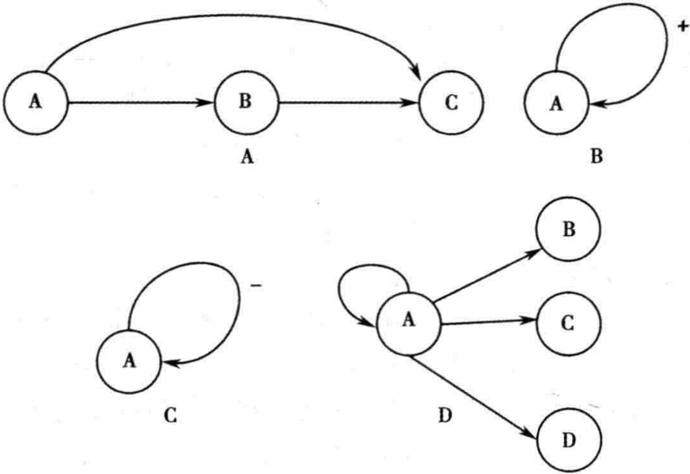


图8-8 网络模序示意图

网络模序结构代表了特定的转录调控机制，对这些模序的研究能够帮助人们了解生物过程的控制机制，因此研究者们开发了网络模序搜索算法来实现在网络中寻找与模序结构同构的子网过程，以发现网络模序。基本的搜索方法是首先定义包含k个节点子网模式；然后搜索网络内全部C_N^k个包含k个节点的节点子集（N代表网路的节点总数），并记录结构与所搜寻的模式相符的次数；最后，将各个模式在真实网络中出现的次数和在随机网络中所出现的次数进行比较，从而发现真实网络中出现次数远超过随机情况的网络模序。由于搜索算法非常耗时，目前网络模序的搜索往往只针对一些较小的子网模式来进行分析。

Mfinder 和MAVisto是两款搜索网络模序的软件，mfinder（http://www.weizmann.ac.il/mcb/UriAlon/groupNetworkMotifSW.html）需要通过命令行的形式进行操作，而MAVisto（http://mavisto.ipk-gatersleben.de/）则包含了一个图形界面。两款软件均可以设定特定的网络模序规模并设计随机扰动以获取相应模序出现频率的显著性。

四、网络的功能模块识别

细胞内的分子通常以模块化的形式行使功能。虽然网络模块（network module）没有一种严格的定义，但通常网络模块是指在物理位置和功能上紧密联系的一组节点。如生物网络中的一组生物分子。一个复杂的网络系统中经常包含很多模块，例如在社会网络中，人类往往会以各种兴趣、爱好和关系等结成各种团体。在人类的工业化生产中，也往往有意识地采用模块化设计。小到移动电话、个人电脑，大到航天器械的设计都采用着模块化的设计提高工程效率和稳定性。生物学的网络系统也包含各种模块化现象。例如蛋白质往往结合成复合物来行使生物学功能，而蛋白质与核酸分子所组成的复合物在从核酸合成到蛋白质降解的生物基本功能中都发挥了重要的作用。在生物应激反应过程中，共同调控的生物分子也协同完成使生物体适应内外环境变化的生物功能。这一部分，我们将介绍从网络中发现模块的方法和衡量网络模块化程度的方法。依赖于网络研究领域，网络模块识别一般也可以称为网络聚类和图划分。在本章中，我们将不区分这些概念。

（一）连通组分模块

网络中如果两个节点间能够由一条路径连接，则称这两个节点是连通的。所有能够彼此连通的节点和它们之间的边构成了一个连通组分。计算网络的所有连通组分，即连通子图。每个连通组分形成一个连通组分（connected components）模块。例如，对于图8-9所示网络来说，有两个连通组分模块。这是最简单的模块识别方法，一般用于其他识别模块方法的初始化阶段。该方法有较大的缺陷，如果节点连通性较好，形成的模块的规模将非常大。连通组分还有两个的扩展版本强连通组分和双连通组分。强连通组分（strongly connected component）指有向网络中两个节点从两个方向上都可通达。双连通组分（biconnected component）在组分中的结果有两个非交叠的路径。

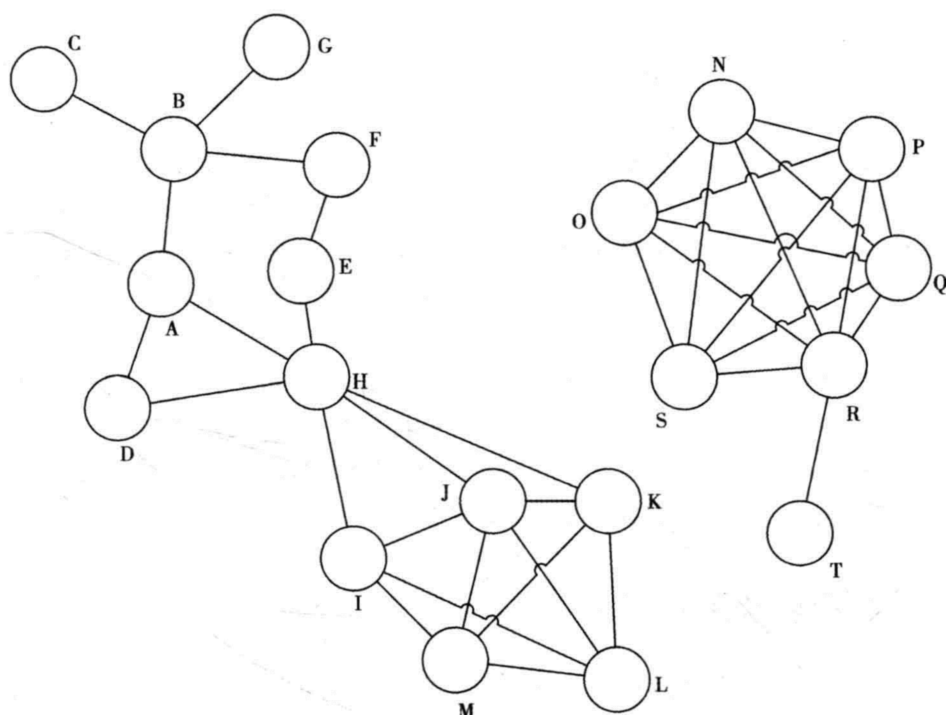


图8-9 连通组分模块示意图

(二) 基于hub的模块

一个基于hub的模块(Hub-based module)包含一个中心hub(度高的节点)和与它距离小于等于d的那些节点。在蛋白质网络中的hub与细胞致死性(lethality)有关,并且与相同的hub连接的蛋白质一般具有相似的功能。基于hub的模块具体识别步骤如下:

1. 计算网络中的每个节点的度。
2. 定义度高于指定阈值(如:大于10)的节点为hub节点。
3. 每个hub和与它距离小于等于d的节点形成一个模块。

对于图8-9所示网络来说,如果设置度为6的节点为hub节点,d设置为1。那么网络将产生两个基于hub的模块M_1={H, A, E, D, I, J, K}和M_2={R, S, O, N, P, Q, T}。

(三) 完全图模块

完全图模块(cliques module)是每对节点都直接连接的图。对于图8-9所示网络,{R, S, O, N, P, Q}是一个由6个节点组成的完全图模块,而{J, K, M, L}是一个4节点的完全图模块。在蛋白质网络中的完全图模块经常对应蛋白质混合物和共同的功能。这种模块也反映了共表达基因的簇。完全图模块在实际应用中可能过于严格。例如,一个蛋白质可能与蛋白质混合物中的多个蛋白质互助而不一定与所有蛋白质互作。而这个蛋白质和与其相互作用的混合物往往行使相似功能。另外,检测技术的缺陷可能导致失去蛋白质混合物中的个别蛋白与其他蛋白的互作。总之,完全图模块有时可能过于严格。因此,palla等人提出全连接集搜索方法(clique percolation method, CPM)来改善过于严格的完全图模块识别。算法如下:

1. 寻找网络中所有的完全图,这将形成多个不同节点数k的clique,即k-cliques(k=3,4, …)。
2. 合并所有共享k-1个节点的k-clique,合并后的子图形成模块,也称为k完全图社区

（k–clique–community）。

如图8–9所示网络，{I，J，K，L，M}是由{I，J，L，M}和{J，K，L，M}两个4–clique形成的4完全图社区模块。我们可知{I，J，K，L，M}并不是完全图，但节点连接非常紧密。palla等人开发了该方法的软件CFinder。它能实现网络密集集团模块搜索和可视化分析。

（四）基于介数的模块识别

Girvan和Newman首次发现边介数对于识别模块非常有效。介数大的边往往是两个模块交互的必经之路。因此，删除介数大的边，将倾向于识别那些功能相对集中的模块。我们将介绍两种基于介数的模块识别方法。介数中心性聚类（betweenness centrality clustering，BCC）和介数共发生分裂（betweenness commonality decomposition，BCD）聚类。BCC也称为GN算法，由Girvan和Newman开发而得名。算法如下：

1. 计算在网络中的所有边的介数。
2. 删除最高介数的边。
3. 重新计算网络中的所有边的介数。
4. 重复步骤2、3直到没有任何边存在。

算法结果将产生层次聚类结构来表示模块，因此该方法属于分裂的层次聚类。因为这个算法每次需要重新计算删除边后的介数，因此，该算法较为耗时。

介数共发生分裂聚类是介数中心性聚类的改进版本。该方法引入共发生性测度来加强连接紧密的蛋白质的聚类，该方法更适合蛋白质网络中寻找功能模块。BCD算法如下：

1. 计算在网络中的所有边的共发生性（C）。边共发生性（commonality）衡量一个边对应的两个端点共享的邻居高于随机发生情况的程度。

$$C = \frac{k+1}{\sqrt{n \times m}} \tag{8-8}$$

n, m代表两个端点的度，k代表共享的邻居数。

2. 计算在网络中的所有边的介数（B）。
3. 删除B/C比值最大的边。
4. 重复步骤2、3直到没有任何边存在。

（五）最大化模块化测度的聚类

一个好的模块划分方案得到的结果应该使得模块内的边更多而模块间的边更少。如果最小化模块间的连接（或最大化模块内的连接），那么最优的划分方案是形成一个单一模块，那样模块间没有任何连接。模块化测度（modularity measure）能够解决这个问题。对于一个网络，如果给定一个划分成模块的方案，模块化M定义为：

$$M \equiv \sum_{s=1}^{N_M}\left[\frac{l_s}{L} - \left(\frac{d_s}{2L}\right)^2\right] \tag{8-9}$$

N_M是模块的数量，L表示在网络中边的数量，l_s代表在模块s中的边数量，d_s代表在模块中所有节点的度的和。

许多算法使用该测度来估计模块识别方法的效果。既然高模块化测度值代表划分方案

高的网络模块化程度。Clauset等人利用最大化模块化测度的策略开发了贪心方法。从一些初始节点出发，迭代的试探加入邻近的节点和边。加入的边保证使模块化M值始终增加，直到值无法继续增加时得到最佳的模块识别结果。Blondel等人开发了局部最优方法，该方法计算初始划分后，将每个划分后的簇当成新的更小网络中的节点，然后在这样更小的网络中寻找保证模块化M值增加的划分。直到M不增加算法停止。该方法速度快，聚类效果也往往比前述的贪心方法好。Guimera和Amaral等人提出模拟退火算法寻找使M最大化的划分。枚举所有划分非常耗时，该方法使用模拟退火算法进行快速搜索，实现寻找目标函数最优的模块划分。因为该方法是全局优化方法，因此得到的效果往往比上面的局部最优方法好。

除了以上所述的模块识别方法外，还存在大量的方法用于网络的模块识别。如原始用于基因表达谱的凝聚层次聚类方法可以用于网络的模块识别，这需要在聚类前先把网络表示为邻接矩阵形式。它还适用于二分网络的聚类。二分网络只需使用双向聚类方法即可实现二分网络模块识别。社会网络的k-clique算法也可以用于网络的模块识别。一个社会网络的k-clique为网络中任意节点之间距离小于等于k的子图，相比完全图模块，社会网络的k-clique模块更强调节点间距离远近关系。

五、网络分析软件

(一) Cytoscape软件

Cytoscape(http: //www.cytoscape.org)是最强大的图形化可视化、编辑和分析生物学网络的软件，界面如图8-10所示。该软件支持多种网络输入格式，也可以使用软件提供的编辑器直接构建新的网络。Cytoscape还能够为网络添加丰富的注释信息，也可以方便地加载自身以及第三方开发的大量功能插件。由于许多网络研究人员向Cytoscape官方网站提供大量的网络分析和可视化插件，使得Cytoscape包含了许多功能强大和特异性的插件，例如生物学网络的各种最新的网路分析方法插件。

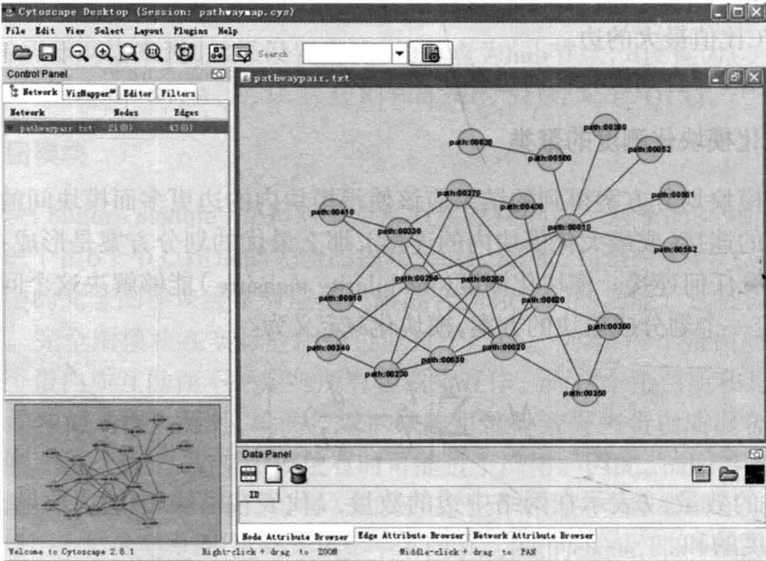


图8-10 Cytoscape软件界面

(二) 基于R的网络分析软件：RBGL和igraph包

R语言系统（http: //www.r-project.org）最早是由Robert Gentleman和Ross Ihaka开始编制，系统界面如图8-11所示。自从基于R的bioconductor项目启动以来，R已经成为生物数据处理和分析最强大的工具之一，目前有许多基于R的网络分析程序包。最强大的包是igraph（http: //igraph.sourceforge.net/）和bioconductor的RBGL包（http: //www.bioconductor.org/）。他们提供了大量的函数用于网路构建、不同形式的输出。网络分析功能非常强大，包含了基本的度、介数等数十种网络拓扑属性测度。也包含评估网络，如power-law分布和寻找各种网络模块的方法，如GN、基于介数、最大化模块化测度的聚类算法。使用R来进行网路分析的最大特点当属用户可以方便地自定义开发新的网络分析方法和改变原有方法。因此，非常适合有一定编程基础的网络分析用户使用。

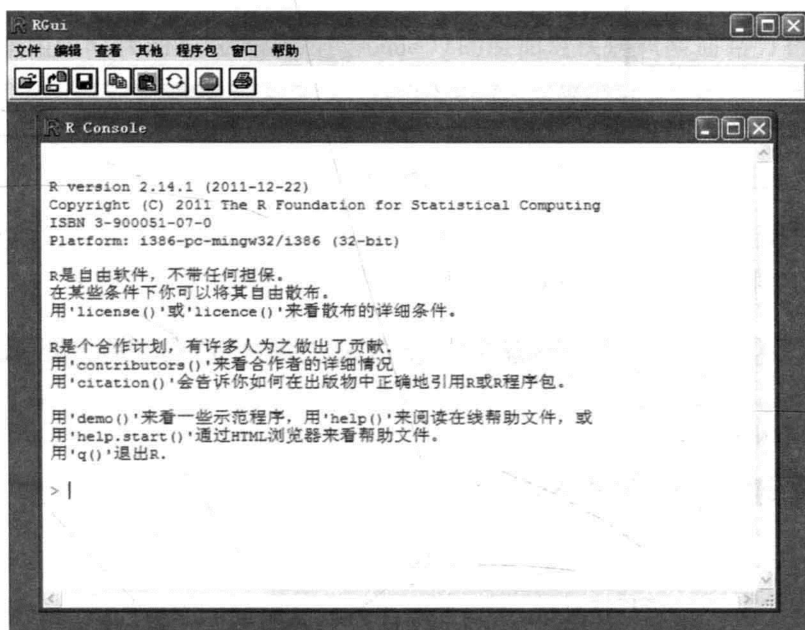


图8-11 R系统的界面

(三) CFinder软件

CFinder（http: //cfinder.org）提供全连接集搜索方法和可视化分析，界面如图8-12所示。完全图模块在实际应用中可能过于严格，例如，检测技术的缺陷可能导致失去蛋白质混合物中的个别蛋白与其他蛋白的互作。CFinder提供全连接集搜索方法来改善过于严格的完全图模块识别，并且可以获得交叠的模块，这更符合生物学的模块含义。

(四) GraphWeb网站平台

GraphWeb（http: //biit.cs.ut.ee/graphweb/）是一个基于网页形式的生物学网络分析和功能模块识别工具，界面如图8-13所示。该工具提供整合异质性数据和多物种数据来构建有向和无向、加权和无权网络的方法。更具特色的是，该工具提供了多种识别网络的功能模块的方法。包含了发现连通组分模块、基于hub的模块、完全图模块和MCL模块的方法。也提

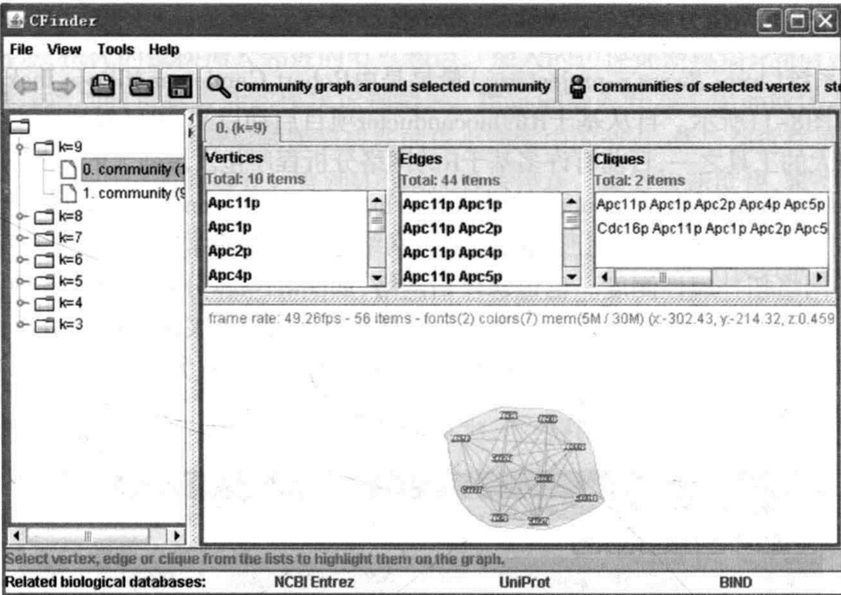


图8-12 CFinder软件的界面

供解释这些模块的生物含义的策略。模块中的基因能够自动的注释到GO或KEGG等数据库来发现模块的生物学功能。GraphWeb基于网页形式,用户操作非常方便。

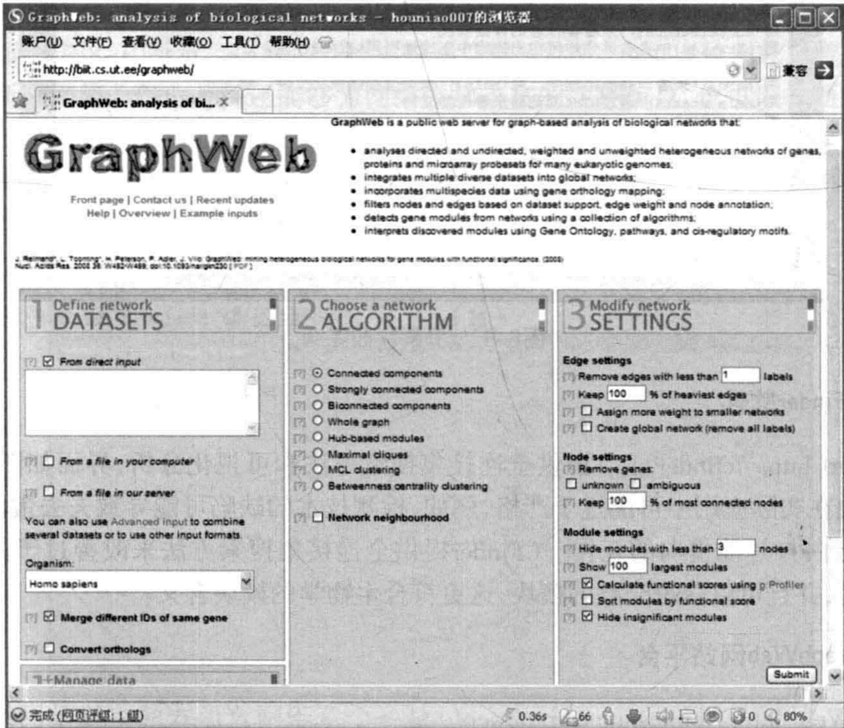


图8-13 GraphWeb网站界面

第四节 通路的网络分析方法

Section 4 Network-based Methods of Pathway Analysis

随着后基因组时代的来临，从组学（"omic"）的层面对疾病风险通路分析已成为一种必然而又合理的趋势。高通量生物芯片、大规模基因突变检测等技术的发展产生了大规模的，几乎涵盖了各种常见疾病的基因数据。研究者们利用这些数据在分析疾病相关代谢通路、信号通路方面取得了很好的研究效果。其中最常用的通路分析是识别与兴趣问题相关的通路。一个有效的方法是使用前面章节介绍的基因集合富集分析方法（详见第三章）。经典的方式是分析一组兴趣基因列表（如：差异表达基因集）在各个通路上是否过出现，可以使用超几何检验（hypergeometric test）和Fisher精确检验等统计学方法识别显著富集的通路。对于像基因表达谱等全基因组检测得到的数据，还可以使用GSEA（gene set enrichment analysis）方法简化感兴趣基因列表的获得过程，避免兴趣基因集合选择过程的偏好性。然而，这些方法的设计思想都是把通路内的基因简化成集合，忽略通路内基因间已知的相互作用关系。通路数据库中存储的通路信息区分于GO数据库的信息的最大特点就是通路数据具有精确的内部分子的相互作用关系，即通路结构信息。这种忽略通路内基因间已知的相互作用关系的缺陷造成通路分析的精确度明显下降，对通路内已有的先验互作关系也造成了彻底的浪费。另外通路之间的交互信息也非常重要，尤其复杂疾病发生多是由多个通路的协同作用导致异常所致。下面将介绍几种利用网络分析技术有效挖掘通路中的结构信息、交互作用来进行通路分析的方法。

一、影响分析方法

影响分析（impact analysis）方法由Draghici等人提出。该方法既考虑了经典的统计学分析的通路得分，又考虑了基因表达值定量的变化和这些基因在通路中的位置对通路的影响情况。例如，在胰岛素（insulin）通路中，胰岛素受体（insulin receptor，INSR）处于通路的起始位置，它的损坏将导致整体通路受到影响，丧失正常功能。而在这个通路的下游中的许多基因的损坏对通路没有大的影响。因此，Draghici等人认为基因在通路中的位置非常重要，尤其是处于起始或上游位置的基因。胰岛素受体有多种功能，并参与到多个通路中。虽然在胰岛素通路中INSR起到了关键作用，但在黏着点（adherens junction）通路中，除了胰岛素受体外，还有许多酪氨酸激酶受体替代胰岛素受体对通路的作用。因此，在这个通路中胰岛

素受体的损坏对通路的影响，并不像在胰岛素通路中那样强烈。因此，即使同一个受体，在不同通路中因为通路结构不同，对通路影响力也不同。为了解决这些问题，Draghici等人利用通路结构信息，提出了影响分析测度，如下：

$$(IF)P_i = \log(\frac{1}{p_i}) + \frac{\sum_{g \in P_i} |PF(g)|}{|\overline{\Delta E}| \times N_{de}(p_i)} \tag{8-10}$$

第一项 $\log(\frac{1}{p_i})$ 表示通路P_i的超几何检验显著性值（即，基因富集分析结果）。第二项表示通路P_i内差异基因对该通路的整体影响，该项值利用了通路结构信息进行计算，依赖于基因在通路中的表达值和通路内基因的互作。其中$N_{de}(p_i)$为通路P_i内的差异表达基因数；$|\overline{\Delta E}|$为平均基因表达量；$PF(g)$表示该通路内基因g的影响得分，它由基因g自身的得分和上游的基因影响得分构成，计算公式如下：

$$PF(g) = \Delta E(g) + \sum_{u \in US_g} \beta_{ug} \times \frac{PF(u)}{N_{ds}(u)} \tag{8-11}$$

$\Delta E(g)$为基因g的差异表达量；US_g为基因g在该通路中的所有上游基因；$N_{ds}(u)$为基因u的下游基因数；如果基因u正调控基因g，则$\beta_{ug}=1$，否则$\beta_{ug}=-1$。从$PF(g)$计算公式可知，如果一个差异基因出现在通路的上游，那么它将对下游的许多基因影响得分具有贡献。差异程度大，贡献也越大。而出现在下游的差异基因，贡献力有限。因此影响分析方法更加强调通路上游基因的影响作用。影响分析方法的第二项与谷歌（Google）网站的页面排序方法类似。如果一个网页有许多网页指向它，那么这个网页是重要的，而对于基因来说是，一个基因能够影响通路下游的许多基因，那么这个基因在通路中更加重要。

Draghici等人将该方法用于肺癌和乳腺癌的风险通路识别，取得了非常好的效果。通过与超几何和GSEA方法比较，发现该方法能够有效识别与疾病相关的通路。尤其是疾病风险基因在通路中分布数量少，但在通路的起始位置起到关键作用的那些通路。由于该方法针对信号通路的特点开发，因此非常适用于信号通路的识别，对代谢通路的识别性能可能效果不佳。

二、潜能通路识别分析

Pam等人开发的潜能通路识别分析（latent pathway identification analysis，LPIA）方法强调通路间的交互重要性。例如，癌症的发生和转移往往与多个通路的交互作用联合导致的异常密不可分。而且，与更多相关的异常通路密切的通路在癌症的发生发展中更为重要。因此，这样的通路更应该被方法有效加以识别。为了实现这一功能，首先应该构建通路和通路之间的交互网络。Pam等人利用每个通路中的基因集合获得与该通路相关的功能，如GO功能。然后根据共享功能的程度，将通路与通路联系起来，形成一个网络，如图8-14A所示。通路是网络中的节点，通路与通路之间的边代表共功能。边具有权值，代表共功能的程度。这个程度根据通路中共享功能的基因数和基因差异表达量来获得。具体计算方法如下：

$$w_{GP} = \left|\frac{G \cap P}{G \cup P}\right| \times med\left\{DE_x : x \in G \cap P\right\} \quad (8\text{-}12)$$

P表示通路，G表示功能，也就是GO中的一个功能项（term）。DE代表基因的差异表达值。Med代表中位值。$G \cap P$为通路P中具有功能G的基因，$G \cup P$为通路P和功能G中所有的基因。如果一个通路包含更多与对应功能相关的基因，并且这些基因的差异表达量更大，那么w_{GP}的值越高。因为两个通路有可能共享多个功能，因此，测度A_{ij}整合了它们的所有权值w_{GP}作为衡量两个通路的交互得分。如下：

$$A_{ij} = \sum_{k=1}^{G} w_{G_k P_i} \times w_{G_k P_j} \quad (8\text{-}13)$$

每两个通路计算得分后，将获得通路的边加权网络，如图8-14B所示。对该网络使用随机游走方法，可以计算每个通路的交互重要性，并对所有通路根据它们的交互重要性与随机情况比较，获得每个通路的交互显著性。交互显著性越强代表这个通路更可能与该疾病相关。例如，按照该方法原理，图8-14B中的通路A将与该疾病最为显著相关。

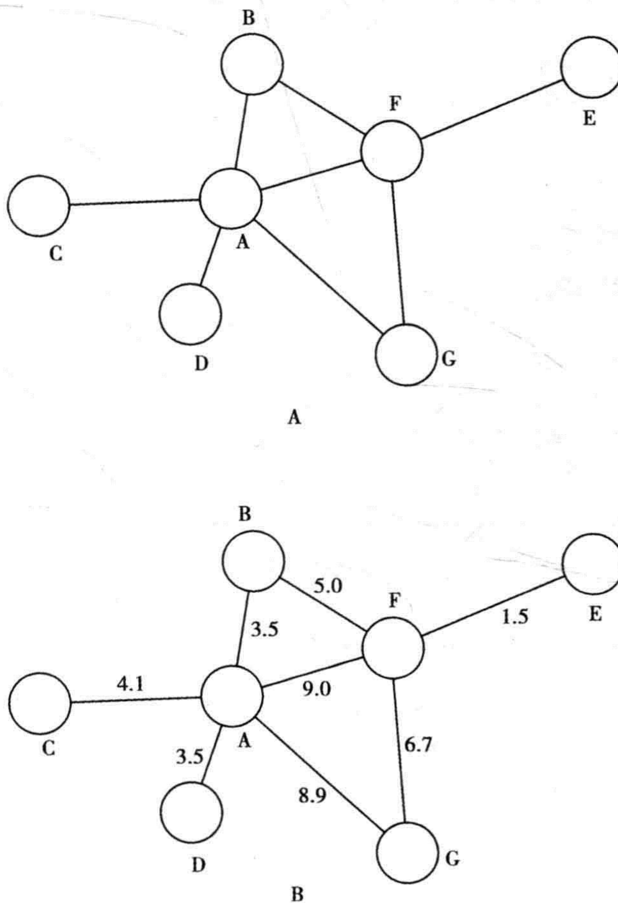


图8-14 LPIA方法构建的通路网络示意图
A.通路为节点的网络；B.边具有权重的网络；
即对A网络中的边赋予交互程度

三、通路分析软件

(一) DAVID软件

DAVID(http: //david.abcc.ncifcrf.gov/)是一款强大的对基因进行通路和功能注释的网站工具,界面如图8-15所示。在网站中输入一个基因集合后,人们能够识别显著富集的功能和通路。改进的Fisher精确检验被用于通路的识别。该工具提供了Biocarta和KEGG通路的可视化图,可以对注释基因在通路中进行表示。而且基因的各种ID支持和转化功能非常强大。

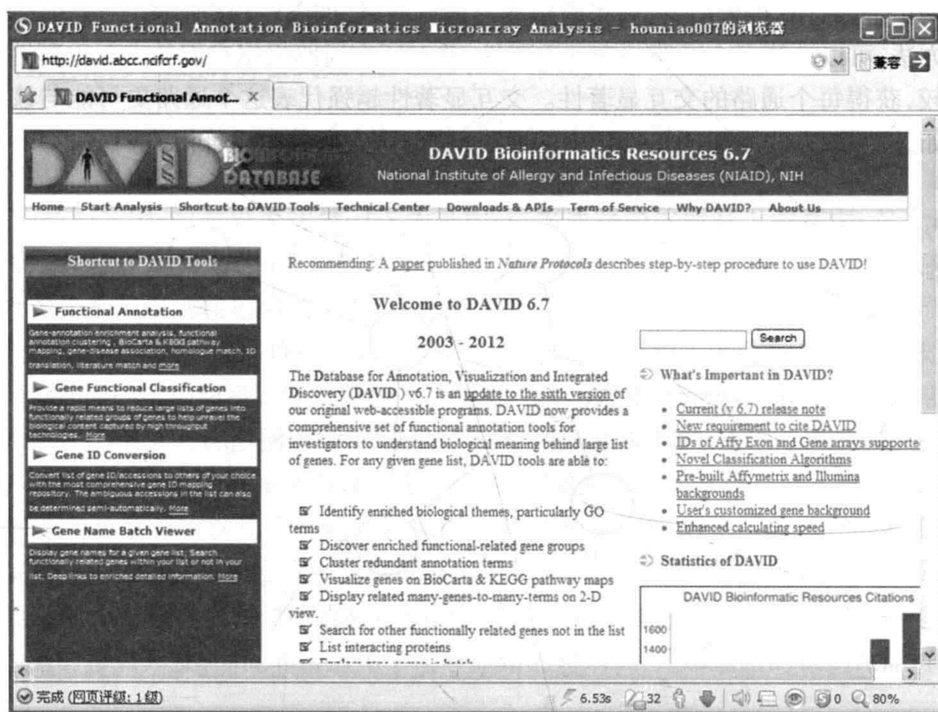


图8-15 DAVID网站界面

(二) 基于R的网络分析软件: iSubpathwayMiner包

iSubpathwayMienr (http: //cran.r-project.org/package= iSubpathway Miner) 是基于RBGL, igraph包开发的专为KEGG代谢通路和信号通路分析程序包,工作界面如图8-16所示。该软件包含了几十种通路图重构方案。多种风险通路和子通路识别方法,通路的拓扑分析方法。通路图可以在R中显示,也可以输出为cytoscape等软件接受的格式,或自动转入KEGG网站进行可视化。

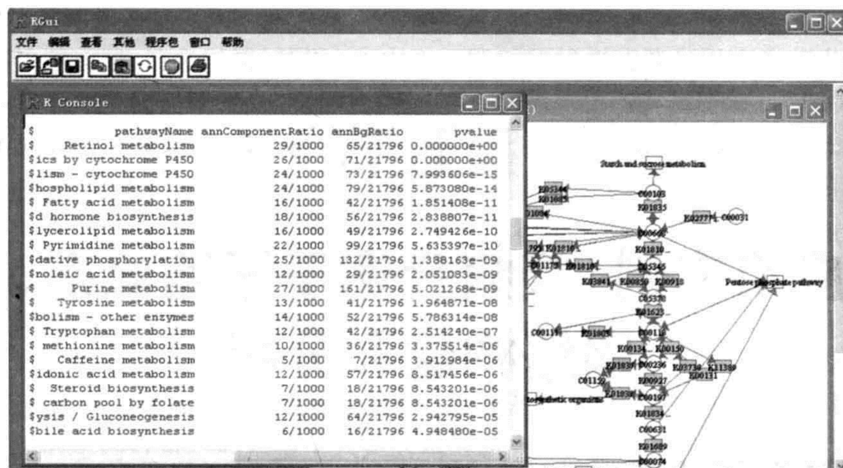


图8-16　iSubpathwayMiner工作界面

(三) pathway–express软件

pathway–express（http://vortex.cs.wayne.edu/projects.htm#Pathway–Express）是影响分析方法对应的平台，界面如图8–17所示。该平台接收用户输入的兴趣基因和量化的值，如由表达谱找到的疾病差异基因和差异倍数（fold change）值。然后通过计算每个通路的影响分析得分，最终返回识别风险通路的结果。影响分析方法既考虑了经典的统计学分析的通路得分，又考虑了基因表达值定量的变化和这些基因在通路中的位置对通路的影响情况。Draghici等人认为基因在通路中的位置是重要的，尤其是处于起始位置的基因。该方法用于肺癌和乳腺癌的风险通路识别，取得了非常好的效果。

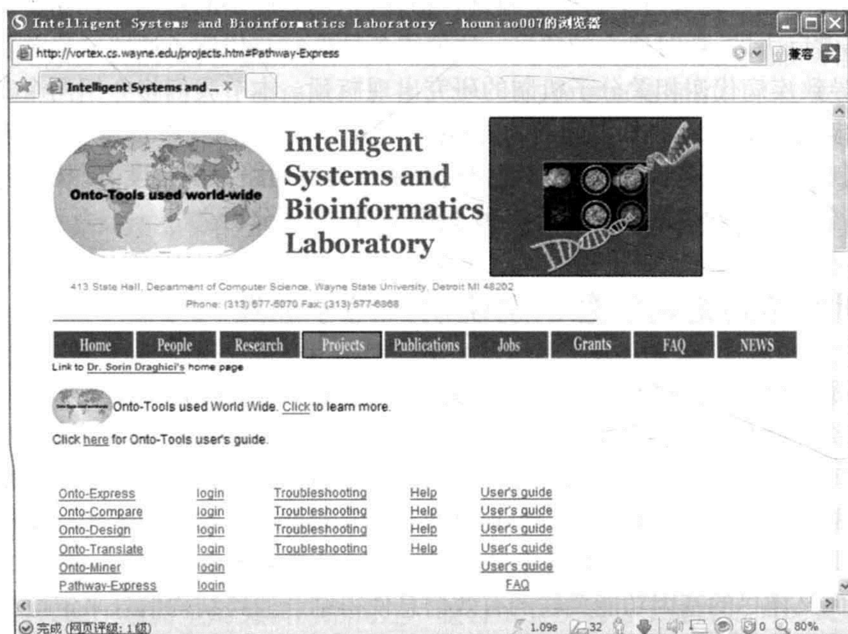


图8–17　pathway–express网站界面

第五节

应用实例：疾病代谢子通路识别、网络构建和分析

Section 5 Application Example: Identification, Network Construction and Analysis of Disease Metabolic Subpathways

疾病的发生和发展往往与代谢通路的异常变化密切相关。随着代谢通路网络数据越来越完善，探索各种疾病尤其是复杂疾病与代谢通路的内在更精细的关联机制更成为后基因时代的一大挑战。因为疾病的发生往往与代谢通路强烈的局部功能和生物学过程混乱密切相关。因此，识别疾病显著相关的代谢子通路区域，能够更加精确地定位疾病相关的代谢局部功能区域和模块。传统上，研究机构一般利用生物学实验技术来定位疾病相关的代谢子通路。然而，这种精细的代谢子通路识别实验一直以来都是生物学和医学领域的难点。由于通路自身分子机制的高度复杂性，使得利用生物学实验方式进行代谢子通路识别相关研究整体进展十分缓慢，仅仅集中在研究个别热点疾病的潜在热点致病代谢子通路上。即使当前的技术能够逐一地通过生物学实验进行筛查，识别如此多的疾病与代谢子通路关系显然是一项非常巨大的、耗时费力的项目。这使得各种疾病与代谢子通路全局关联关系无法清晰呈现，导致疾病代谢相关分子机制的研究出现瓶颈。本节我们将介绍两个实例来演示利用通路网路分析技术改善如上问题的方案：①使用社会网络的k-clique方法和利用通路结构信息识别疾病风险子通路来改善通路识别效果及提高通路识别精度；②构建各种疾病与代谢子通路全局关联网络来系统分析疾病通路机制的案例分析。

一、利用通路结构信息识别疾病风险子通路

代谢通路整体上是一个复杂的网络，通路内几乎所有的组分（酶、化合物等）彼此之间通过数步的级联生化反应相关联。一个酶基因的表达变化（如编码该酶的基因的突变）可能影响网络内的另一个酶基因的变化。然而，这种影响有强弱之分，往往通路内彼此邻接越近的酶之间，相互影响的程度越大，它们也更倾向于具有相似的生物学功能和行使相同的生物学过程。因此，精细定位和识别疾病相关代谢子通路局部区域意义重大。通路结构信息中隐含了大量而又详尽的基因功能关联的有效信息使得结合通路结构信息来定位和识别子通路是十分有应用价值的。本实例介绍使用k-clique方法和利用通路结构信息，将代谢通路划分成子通路，并利用超几何检验方法对评估和识别疾病风险子通路的案例。k-clique代谢子

通路识别方法具体如下：

1. 重构代谢通路，使之以酶（基因）为中心 即转化代谢通路为以酶为节点，酶和酶之间如果共享至少一个代谢子，那么它们之间连接一条边。如图8-18所示。A图为代谢通路的生化反应图，经过转化后将得到酶为节点的图，即B图。

2. 接下来，使用图划分算法，将通路划分成子通路（网络模块） 社会网络的k-clique算法是一个理想的选择。一个k-clique为图中任意节点之间距离小于等于k的子图（图8-18C）。利用k-clique算法能够将每个通路图都划分成子图，子图对应的通路部分称之为子通路。

3. 注释疾病差异基因到相应每个代谢子通路中。

4. 最后利用超几何富集分析技术识别注释到相应每个代谢子通路的富集显著性。通过基因-酶对应关系将基因集合注释到子通路中。对于基因集合注释到的代谢子通路，统计卷入这个子通路的基因数量。如果整个人类的基因组有m个基因，而这些基因落入子通路的基因数为t。如果提交的基因集合的基因数为n，而这些基因注释到子通路的基因数为r。则可以通过超几何检验（详见第三章内容）计算该子通路的统计学显著性p值。

我们将k-clique代谢子通路识别方法应用到肺癌的表达谱数据，令人满意地识别出与肺癌发生和发展高度相关的具有高度生物学显著性的通路。该方法能够精细定位子通路区域的特点能够有效的细化通路识别，并挖掘出通路整体异常不明显、但局部区域异常显著的通路。社会网络的k-clique方法更强调节点距离远近关系来识别网络模块的特点，使之更倾向于识别具有相似的生物学功能和行使相同的生物学过程子通路模块。该方法的实现提供在iSubpathwayMienr（http://cran.r-project.org/package= iSubpathwayMiner）包，它提供了该方法的使用及相关程序代码。

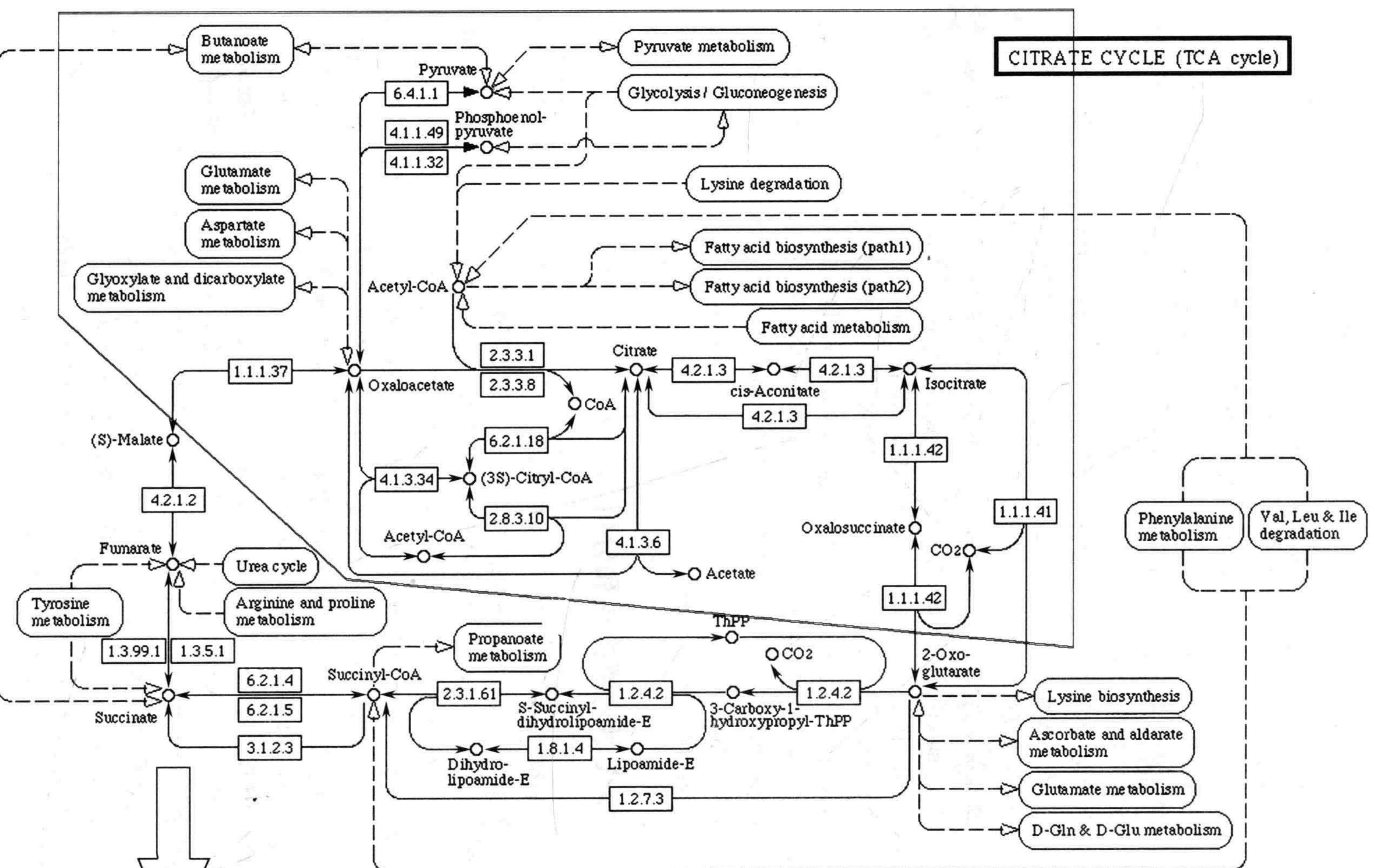
二、疾病代谢网络构建和分析

相似的疾病很可能具有相似的发生发展机制。一个罕见癌症的发生机制可能与常见癌症的发生机制相关，例如相似代谢通路的异常。进一步，如果更精细识别疾病通路的异常区域，即子通路，并能够构建所有已知疾病与子通路的全局关联关系，那么对于疾病通路的研究意义重大。当前的生物实验技术显然无法逐一进行实验筛查。因为识别如此多的疾病与代谢子通路关系是一项非常巨大的、耗时费力的项目。这使得各种疾病与代谢子通路全局关联关系无法清晰呈现，导致疾病代谢相关分子机制的研究出现瓶颈。本实例利用实例一的子通路识别方法识别每个疾病的风险代谢子通路，从而构建所有疾病与代谢子通的全局关联网络，然后利用网络分析方法分析网络。构建过程如图8-19所示，具体过程如下：

1. 获取和处理疾病-基因信息 从genetic association database（GAD）疾病基因数据库中获取疾病分类数据、对应的疾病基因数据，并对这些疾病基因关系进行去冗余、统一命名标准及存储格式、合并亚类疾病等数据整理处理工作，得到疾病及基因的统一关系。

2. 疾病代谢子通路识别 使用疾病基因关系，将每种疾病的基因集合输入到k-clique代谢子通路识别方法中，识别出各种疾病对应的风险代谢子通路区域。

3. 疾病-代谢子通路全局关联网络构建 整合所有疾病与子通路之间的关系成为以疾病和子通路为节点，疾病与子通路关联结果为边的二分图网络，随后继承疾病和通路的类别等信息。


CITRATE CYCLE (TCA cycle)
Butanoate metabolism
Pyruvate metabolism
Pyruvate
6.4.1.1
Glycolysis / Gluconeogenesis
4.1.1.49
Phosphoenol-pyruvate
4.1.1.32
Glutamate metabolism
Lysine degradation
Aspartate metabolism
Fatty acid biosynthesis (path1)
Glyoxylate and dicarboxylate metabolism
Acetyl-CoA
Fatty acid biosynthesis (path2)
Fatty acid metabolism
2.3.3.1
Citrate
4.2.1.3
4.2.1.3
1.1.1.37
Oxaloacetate
2.3.3.8
cis-Aconitate
Isocitrate
CoA
4.2.1.3
(S)-Malate
6.2.1.18
1.1.1.42
4.1.3.34
(3S)-Citryl-CoA
4.2.1.2
2.8.3.10
1.1.1.41
Oxalosuccinate
Phenylalanine metabolism
Val, Leu & Ile degradation
Acetyl-CoA
4.1.3.6
CO2
Fumarate
Urea cycle
Acetate
Tyrosine metabolism
Arginine and proline metabolism
1.1.1.42
ThPP
1.3.99.1
1.3.5.1
Propanoate metabolism
CO2
2-Oxo-glutarate
Succinyl-CoA
6.2.1.4
2.3.1.61
1.2.4.2
1.2.4.2
Lysine biosynthesis
Succinate
6.2.1.5
S-Succinyl-dihydrolipoamide-E
3-Carboxy-1-hydroxypropyl-ThPP
Ascorbate and aldarate metabolism
3.1.2.3
1.8.1.4
Dihydro-lipoamide-E
Lipoamide-E
Glutamate metabolism
1.2.7.3
D-Gln & D-Glu metabolism
A

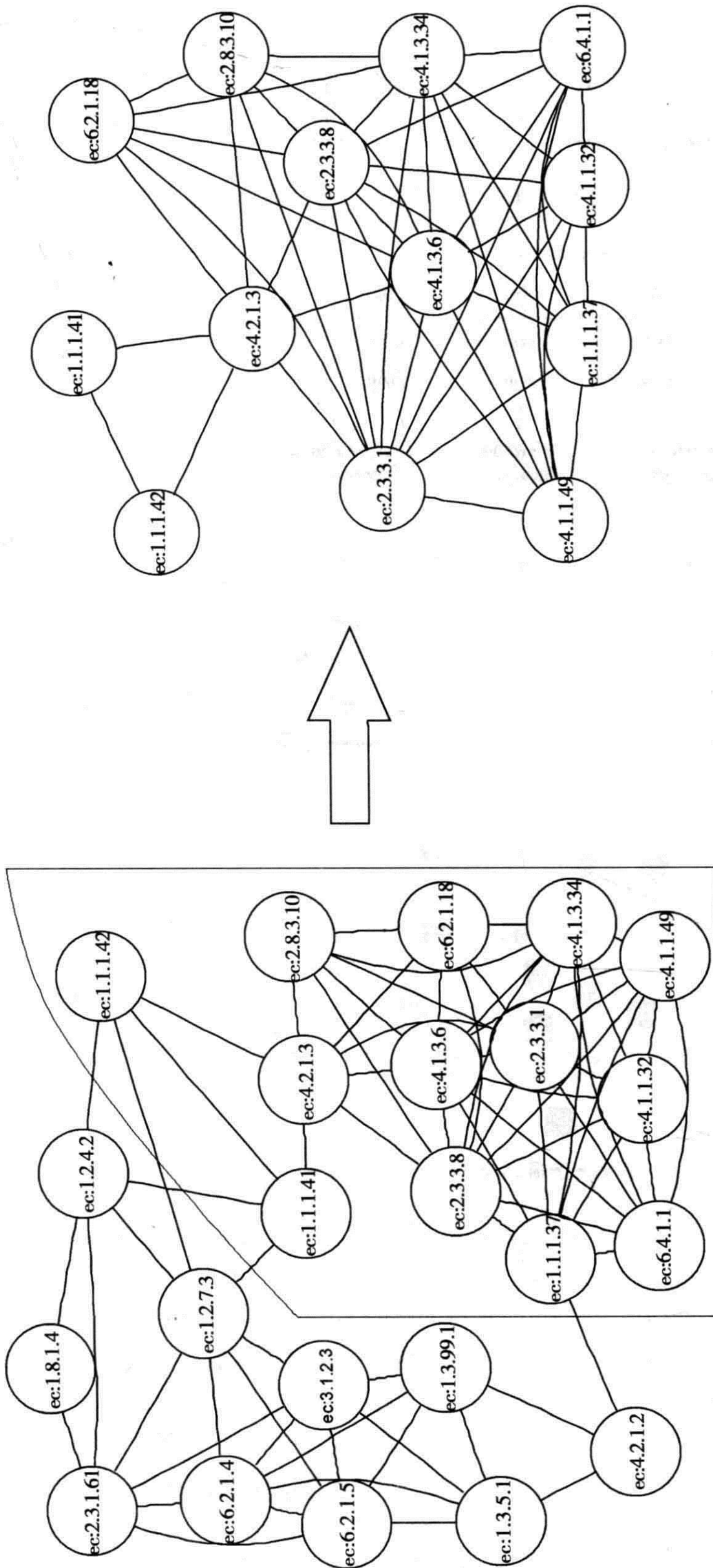


图8-18　A TCA cycle通路以化合物为中心的表示方式；B 通路转化为以酶为节点的表示方式；C 利用3-clique算法得到的一个子通路

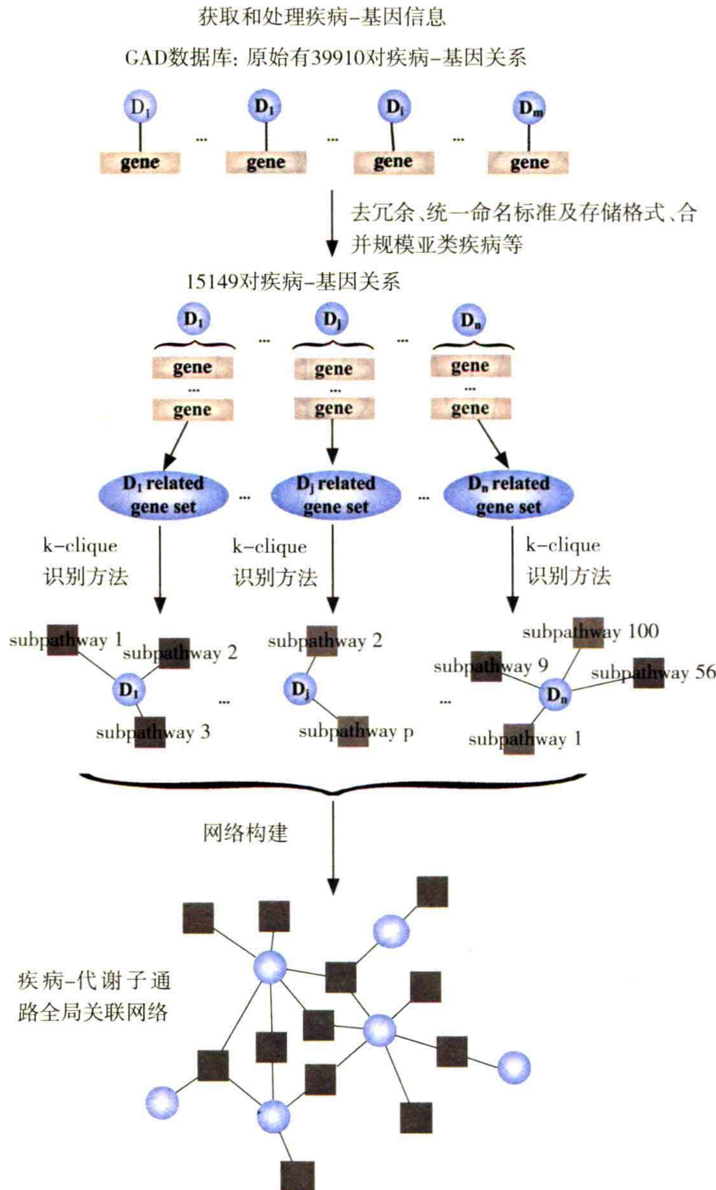
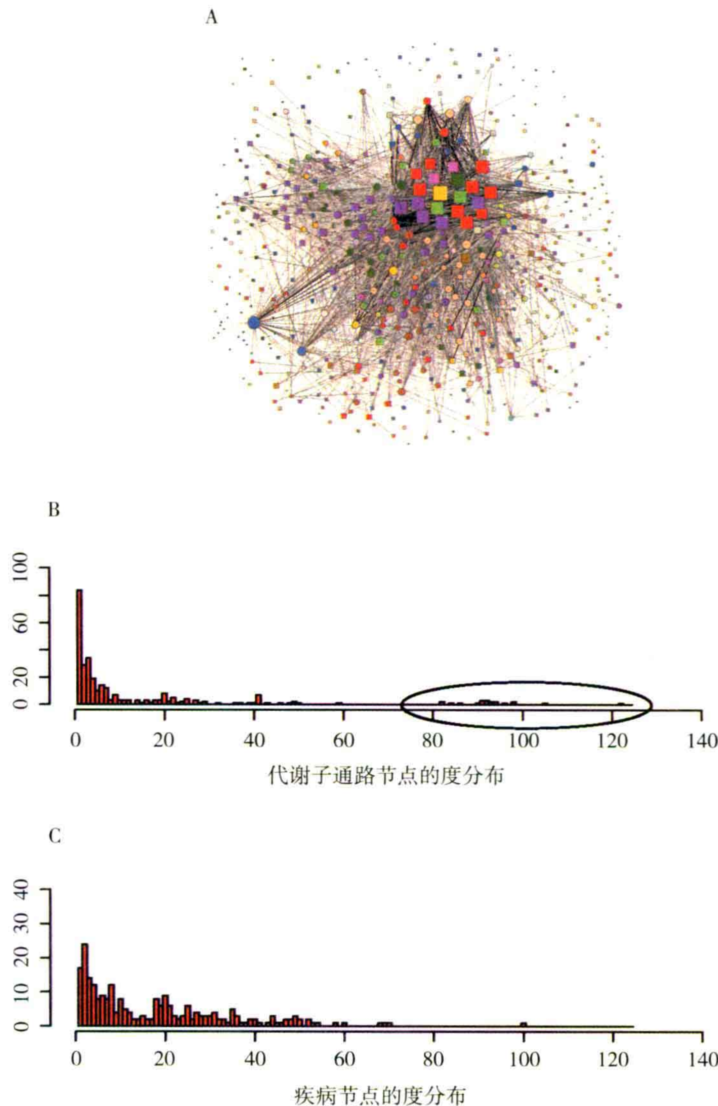


图8-19 疾病-子通路全局关联网络构建过程

通过上面构建网络的过程，最终能够构建疾病-代谢子通路全局关联网络。结果如图8-20A所示：方块节点表示代谢子通路，圆表示疾病；颜色表示通路所属类别和疾病所属类别。构建完疾病-代谢子通路全局相关网络后，我们可以进行网络的基本属性拓扑分析。度分布特性分析是最常见的分析，即计算网络中每个节点的度，然后统计每个度出现的次数或频率。因为这里构建的网络是二部网络，所以我们分别计算疾病和子通路节点的度分布特性更加合理和符合生物学本质。如图8-20B、C所示，疾病节点和代谢子通路节点的度分布都大致服从幂律分布。如对于代谢子通路来说，大部分的子通路仅仅与少数疾病的发生、发展有关。仅仅很少的一批子通路与大多数的疾病相关。这些子通路位于图8-20A的网络中心的大方块节点，它们的度如图8-20B的黑圈区域所示。进一步我们发现，这些通路都属于基础类代谢通路，这显示了疾

病的发生发展往往都与基础类代谢的异常有关。接下来我们利用网络聚类方法识别网络的疾病-代谢模块。我们使用双向层次聚类方法对网络进行聚类，来识别网络的模块。如图8-20D所示，结果显示网络具有模块化的倾向，且同类疾病和代谢子通路倾向于在相同或相近的模块。这显示了相似疾病倾向于共享更多相同的代谢子通路。进一步将深入探讨潜在致病代谢子通路内部基因成分与通路致病性强弱分析。疾病-代谢子通路全局关联网络的一大优势是可以从系统的角度分析疾病风险通路中酶基因的变化规律。我们考察了疾病基因、必要基因（essential genes）在子通路中的含量的变化。发现可能呈现不同的趋势。在网络中度与子通路中不同类型基因的含量之间的关系显著相关。当一个代谢子通路中疾病基因含量较多时，这个子通路更倾向于与更多类型的疾病相关，如图8-20E所示。然而，当一个代谢子通路中必要基因含量较多时，这个子通路更倾向于导致更少的疾病发生，如图8-20F所示。这一现象可能反映了大部分疾病的发生不会破坏必要基因丰富的通路。疾病-代谢子通路全局关联网络分析可以从系统角度帮助理解疾病与疾病、疾病与代谢子通路、代谢子通路与代谢子通路之间的关系。



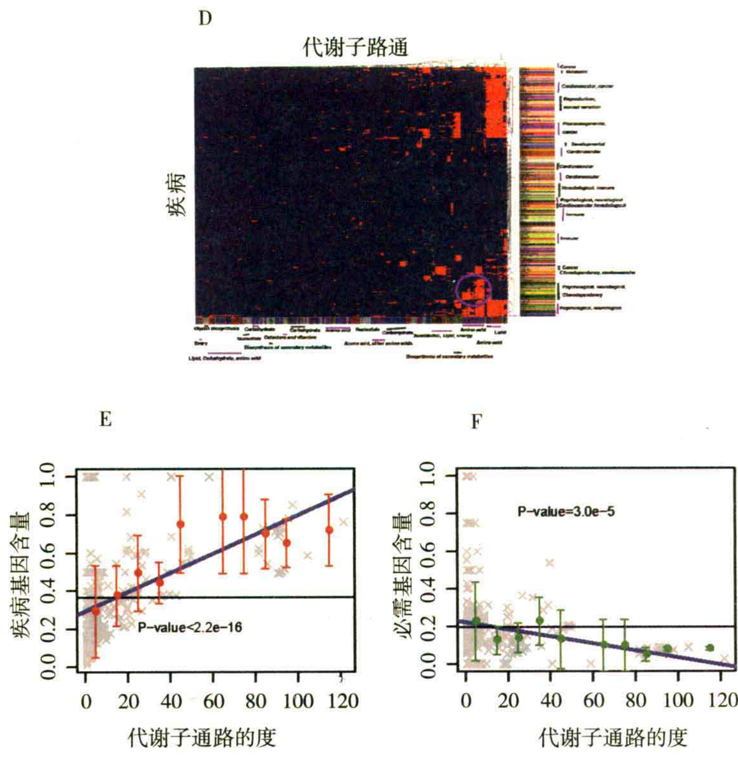


图8-20 A 疾病—子通路全局关联网络

A.节点大小表示节点的度，即一个疾病的发生与多少代谢子通路相关或一个代谢子通路可能导致多少疾病；B.网络中代谢子通路的度分布；C.网络中疾病节点的度分布；D.使用双向聚类方法对网络进行聚类的结果；元素中橙色表示对应的疾病和子通路相关；右侧的颜色条中的颜色表示疾病的类，同颜色的疾病代表他们属于相同的疾病类；E.代谢子通路的度与通路中的疾病基因含量的关系；F.代谢子通路的度与子通路中的必需基因含量的关系

（李春权）

参考文献

1. Barabasi A. L. *Network medicine--from obesity to the "diseasome".* N Engl J Med, 2007. 357(4): 404-407.
2. Barabasi A. L, Oltvai Z. N. *Network biology: understanding the cell's functional organization.* Nat Rev Genet, 2004. 5(2): 101-113.
3. Draghici, S., Khatri, P., Tarca, A. L., et al., *A systems biology approach for pathway level analysis.* Genome Res, 2007. 17(10): 1537-1545.
4. Fang Z, Tian W, Ji H. *A network-based gene-weighting approach for pathway analysis.* Cell Res, 2011. 22(3): 565-580.
5. Girvan M, Newman M. E. *Community structure in social and biological networks.* Proc Natl Acad Sci U S A, 2002. 99(12): 7821-7826.

6. Goh, K. I., Cusick, M. E., Valle, D., et al., *The human disease network.* Proc Natl Acad Sci U S A, 2007. 104(21): 8685–8690.

7. Guimera R, Amaral L. A., *Functional cartography of complex metabolic networks.* Nature, 2005. 433(7028): 895–900.

8. Huang da W, Sherman B. T, Lempicki R. A. *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Res, 2009. 37(1): 1–13.

9. Huang da, W., Sherman, B. T., Tan, Q., et al., *DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists.* Nucleic Acids Res, 2007. 35(Web Server issue): W169–175.

10. Kanehisa, M., Goto, S., Hattori, M., et al., *From genomics to chemical genomics: new developments in KEGG.* Nucleic Acids Res, 2006. 34(Database issue): D354–357.

11. Li, C., Li, X., Miao, Y., et al., *SubpathwayMiner: a software package for flexible identification of pathways.* Nucleic Acids Res, 2009. 37(19): e131.

12. Li, X., Li, C., Shang, D., et al., *The implications of relationships between human diseases and metabolic subpathways.* PLoS One, 2011. 6(6): e21131.

13. Palla, G., Derenyi, I., Farkas, I., et al., *Uncovering the overlapping community structure of complex networks in nature and society.* Nature, 2005. 435(7043): 814–818.

14. Papin, J. A., J. L. Reed, and B. O. Palsson, *Hierarchical thinking in network biology: the unbiased modularization of biochemical networks.* Trends Biochem Sci, 2004. 29(12): 641–647.

15. Pham, L., Christadore, L., Schaus, S., et al., *Network-based prediction for sources of transcriptional dysregulation using latent pathway identification analysis.* Proc Natl Acad Sci U S A, 2011. 108(32): 13347–13352.

16. Reimand, J., Tooming, L., Peterson, H., et al., *GraphWeb: mining heterogeneous biological networks for gene modules with functional significance.* Nucleic Acids Res, 2008. 36(Web Server issue): W452–459.

17. Schreiber F, Schwobbermeyer H. *MAVisto: a tool for the exploration of network motifs.* Bioinformatics, 2005. 21(17): 3572–3574.

18. Shannon, P., Markiel, A., Ozier, O., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. 13(11): 2498–2504.

19. Wang, C., Ding, C., et al., *Consistent dissection of the protein interaction network by combining global and local metrics.* Genome Biol, 2007. 8(12): R271.

20. Yildirim, M. A., Goh, K. I., Cusick, M. E., et al., *Drug-target network.* Nat Biotechnol, 2007. 25(10): 1119–1126.

第九章
CHAPTER 9

复杂疾病的系统遗传学分析

SYSTEM GENETICS ANALYSIS OF COMPLEX DISEASES

疾病是机体在遗传和环境因素共同作用下，机体稳态失衡而发生的异常生命过程，其表现为细胞、组织或器官层面的损害作用。从分子遗传学角度看，致病因素包括遗传突变、DNA损伤和异常修复、调控紊乱、基因表达或蛋白质功能异常等，往往和环境因素直接或间接地发生作用，从而导致机体产生一系列功能、代谢和形态结构的变化，并由此产生相应的症状和体征。

通常我们把疾病发生关联的因素分为内因和外因，内因主要是染色体异常、基因剪接异常、单核苷酸的插入缺失变异、拷贝数变化等、DNA修饰和核小体修饰等遗传和表观遗传变化，这些变化可能直接导致机体功能先天异常，或使机体对外界刺激的敏感性发生变化。外因是诱发疾病出现或易感的多种外界因素，包括感染、损伤、环境、情绪、教育和社会因素等，当具有某种遗传特质的人接触到不相适应的外界因素时，疾病的发病率可能成倍增加。随着现代分子生物学和医学研究的不断发展，尤其是人类基因组计划（human genome project，HGP）的完成和国际人类变异组计划的开展，积累了大量的疾病分子水平知识和相应研究手段，使我们不仅可以深入到高通量分子遗传学层面认识疾病本质，而且还可能利用这些知识探索和创造疾病诊疗新方法、新技术，指导新的药物靶标发现。

第一节

复杂疾病的分子遗传学特征

Section 1 Molecular Genetics Character of Complex Diseases

一、孟德尔遗传病与复杂疾病

目前所知的大部分疾病与遗传因素密切相关，依据疾病与不同遗传因素之间的联系，可以将疾病进行分类：单基因病、多基因病、线粒体病及染色体畸变所引起的疾病。其中，单基因遗传疾病又称孟德尔遗传病，即医学遗传学中通常研究的常染色体显性遗传病、隐性遗传病，及性染色体连锁的遗传病等，由于这些疾病一般发病率极低（群体发病率低于万分之一），且有较强的肢体致残或致死率，也称为罕见疾病（rare disease）。而人类常见疾病（群体发病率较高），如肿瘤、心脑血管疾病、代谢系统疾病、神经系统疾病等，往往不是由单个基因或者单种因素决定的，而是涉及多种基因、环境及遗传等多方面因素，与孟德尔遗传病相比在成因上具有显著的复杂性，因此称为复杂疾病（complex disease）。

为使疾病研究具有系统性和参照性，人们很早就开始疾病分类学研究。最早的疾病分类体系创建于19世纪50年代，并在1893年由国际统计研究所出版了《International List of Causes of Death》。世界卫生组织（WHO）于1948年开始负责ICD（international statistical classification of diseases and related health problems）的编写任务，并加入了发病原因信息。世界卫生大会（WHA）于1967年通过了世界卫生组织对疾病的命名规则，并要求其成员国使用ICD上疾病命名规则对疾病死亡率和发病率进行统计。ICD疾病分类体系按照疾病特征将其分门别类，现有版本（ICD-10）包含15.5万种编码。各个国家分别引进这种疾病分类体系并进行改进，中国根据“ICD-10”颁布了《第二次国家卫生服务调查疾病分类—编码表》对疾病进行了分类，共19类：①传染病；②寄生虫病；③恶性肿瘤；④良性肿瘤；⑤内分泌疾病（营养和代谢疾病及免疫疾病）；⑥血液和造血器官疾病；⑦精神病；⑧神经系统疾病；⑨眼及附器疾病；⑩耳和乳突疾病，⑪循环系统疾病；⑫呼吸系统疾病；⑬消化系统疾病；⑭泌尿生殖系统疾病；⑮妊娠；⑯分娩病及产褥期并发症；⑰皮肤和皮下组织疾病；⑱肌肉、骨骼系统和结缔组织疾病；⑲损伤和中毒。

二、复杂疾病的分子系统特征

与孟德尔遗传病相比，复杂疾病具有四个独特的分子遗传特征。第一，复杂疾病是多基因病（ploygenic disorder）。复杂疾病的发生往往与多个基因的遗传或表达变化存在联系，可

能是分子层面多个基因损伤、变异、失调而累积产生了基因产物异常(表达数量或蛋白质结构)、代谢失调或信号通路异常等,从而导致机体的宏观变化。第二,复杂疾病致病基因具有微效性(minor effect)。孟德尔遗传病的发生往往只与一个或几个基因的变化相关,具有明显的主效基因,而在复杂疾病遗传遗传学研究中,很难发现一个或几个具有明显致病作用的基因,每个基因对于疾病的发生均具有较为温和的作用效果,疾病的发生是一个从量变到质变的过程。第三,复杂疾病具有遗传异质性(heterogeneity)特征。在临床上,遗传异质性是指不同的成因可能导致相同的临床症状,与此类比,在分子遗传学层面,复杂疾病的异质性指的是分子层面的某些并不完全相同的变化累积可能导致同一个疾病的发生,这与人群种族差异、基因功能关联性、环境因素刺激等密切相关。第四,复杂疾病相关基因存在上位效应(epistasis)或相互作用。复杂疾病的发生与众多基因相关,但这些基因之间并不是孤立发生作用的,而存在紧密的调控或互作关系,这种关系可以将作为启动点的几个基因的作用放大到某一生物学过程或生物通路层面,将分子异常引入到宏观机体表现。复杂疾病的分子遗传特征决定了其病因学研究过程的艰巨性。随着基因表达检测技术、SNP检测技术、蛋白质检测技术、表观遗传检测技术等高通量分子标记检测方法的迅速发展,人们已经开始着眼于基因组范围系统地研究复杂疾病的发生过程,从这些方面入手开发快速有效的生物信息学分析工具和方法具有重要的意义。

遗传因素之外,环境因素对于复杂疾病的形成有着非常重要的作用。据世界卫生组织报告,全球超过20%的疾病是与环境暴露直接关联的。每年约有1300万人死亡归因于环境的不适应性。在人类分布最不发达地区,近三分之一疾病可以归因于环境因素。同时,报告还指出,在造成人类死亡率最高的几种疾病中(如心血管疾病、呼吸系统炎症、癌症、慢性阻塞性肺病等),85%以上的疾病受环境因素影响。环境导致人类疾病的发生主要体现在两个方面。首先,某些环境条件可能会诱导基因发生突变或表达变化引发疾病。比如癌基因在通常情况下处于抑制状态,当细胞被紫外线照射或者受到异常环境因素刺激,癌基因就可能从原来的抑制状态变成激活状态,进而使得正常细胞发生癌变转化为癌细胞。其次,人类个体本身的遗传差异在一定程度上决定了人们对环境的适应性差异,即基因型差异影响到对环境改变的敏感性。越来越多的实验证明基因与环境之间的相互作用在复杂疾病的发生或发展过程中起着关键性作用,它们之间的相互作用是极其复杂和非线性的。一个基因在不同的环境中会产生不同甚至是完全相反的表型,因此单纯从遗传角度去研究疾病是不足以全面了解复杂疾病的发生、发展过程的。为了全面、系统地研究环境对于疾病的影响,科学家们开展了环境基因组计划(environment genome project, EGP),识别哪些人类基因能增加对环境相关疾病的个体易感性。

三、重要的复杂疾病数据库

(一) 人类孟德尔遗传在线(OMIM)

MIM(Mendelian inheritance in man)是一个将遗传病进行分类,并与人类基因建立相互联系的疾病研究数据库。它的在线版本是人类孟德尔遗传在线(OMIM, http: //www.ncbi.nlm.nih.gov/omim)。OMIM是目前最权威的人类遗传疾病数据库,为临床医生和科研人员

提供了权威可靠的遗传疾病、表型相关基因或染色体位点信息，有着广泛的应用领域，对于临床医生和科研人员来说是一种重要的网络资源。例如，临床医生可以将患者的临床表型输入到数据库查找相关的疾病信息，又可以针对某些感兴趣的基因或者疾病进行搜索。在OMIM中搜索基因和疾病时，又可以同时查询到基因和疾病相关的信息如基因的序列，染色体位置，以及疾病相关的参考文献等。OMIM提供了友好的使用界面，用户可以通过MIM号（ID）、疾病名、基因名或者疾病的一些表征进行搜索（图9-1）。

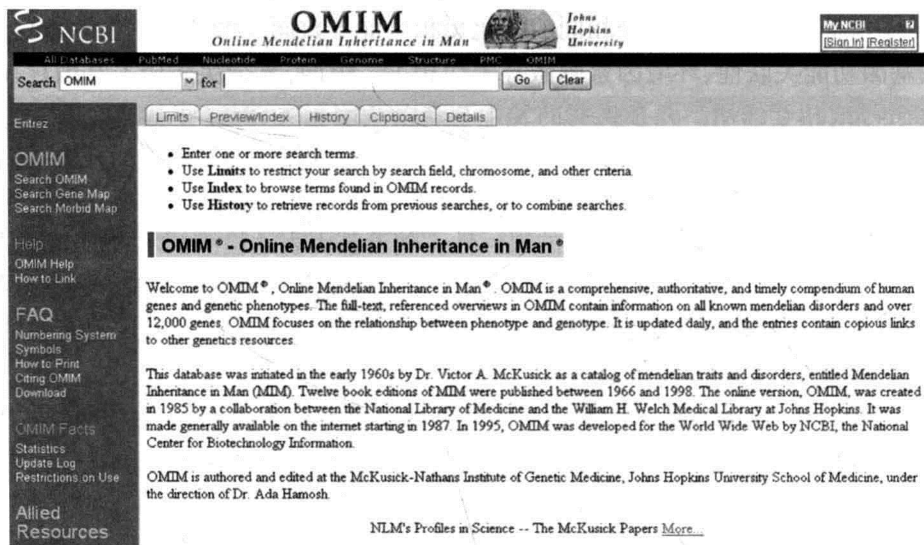


图9-1 OMIM在线搜索界面

在输入搜索关键词并运行后，网站会在搜索结果中列出与搜索记录最相近的20个记录，读者可依照个人习惯更改显示记录的数目。在OMIM数据库中，每一个记录都会有唯一的6位数编码，这种编码可以表示这种遗传病是常染色体显性（隐性）遗传、X连锁还是Y连锁等，详见表9-1。

表9-1 OMIM编码及其代表的数据类型

OMIM编号范围	遗传方式
100000–199999	常染色体显性遗传或表型（于1994年5月15日创建）
200000–299999	常染色体隐性遗传或表型（于1994年5月15日创建）
300000–399999	X连锁位点或表型
400000–499999	Y连锁位点或表型
500000–599999	线粒体位点或表型
600000–	染色体位点或表型（于1994年5月15日创建）

在大部分OMIM编码前会有一些特殊的符号来分别表示不同的含义。其中，"*" 表示本条目为某个基因的系统注释；"#" 表示本条目为某个表型的系统说明；"+" 表示本条目代表基因型与表型之间的关系；"%" 表示某个未知分子机制的疾病表型或相关位点；条目前无

任何标记表示此疾病表型可能存在遗传相关位点，但未经证实；“^”表示该记录已不存在或者被其他记录所代替。此外，OMIM数据库还提供批量下载，用户可以通过FTP登录方式下载OMIM中的全部数据（ftp: //ftp.ncbi.nih.gov/repository/OMIM），包括OMIM全部文本信息文件（omim.txt.Z），OMIM中收录的基因（genemap）、基因说明文件（genemap.key），以及疾病与基因对应关系文件（morbidmap）。

另外，OMIM还提供genemap和morbidmap的网络查询形式。这里以Alzheimer’s Disease（AD）为例，简单介绍一下OMIM数据库的使用。我们在OMIM查询框中输入“Alzheimer’s Disease”就可以在OMIM上得到这种疾病相关信息。读者也可以输入该疾病的简写形式AD，疾病表征（如Senile Dementia）、或者与该疾病相关的基因名（*APOE4*），查看检索结果（因关键词影响，结果略有差别）。

其中每一个记录表示在OMIM中与查询信息相关的内容。另外，我们可以在“Display”中选择查询结果的显示方式、条目数。选择任意一条记录都包含了如下信息：MIM号（ID）、查询疾病的名称（别名），与疾病相关遗传信息的一般性描述，有文献支持的临床表征，生化特征，发病机制，遗传性及诊断，文献支持的基因信息，分子遗传学、群体遗传学等文献支持材料。最后，提供了大部分的研究参考文献。选择页面上的Gene map locus后面的基因区段，会显示出该区段在染色体图的详细信息。主要的内容包括如下几方面的图信息：基因序列信息、表型信息（包括数量性状位点）、OMIM疾病记录、细胞遗传上的基因分布等详细信息。其中部分数据可以下载或查看。

（二）遗传关联数据库（genetic association database，GAD）

GAD是由美国国立卫生研究院（national institutes of health，NIH）的Kevin Becker及其同事于2004年开发维护的数据库，该数据库中存储了大量的人类复杂疾病相关的基因及多态性信息，为研究人员从大量的多态性数据中快速地识别出疾病相关的多态提供了方便。数据库中的信息来源于对目前已有的关联分析结果的搜集和整理，这些信息是以基因为核心的，也就是说，数据库中的每条记录对应的是一个基因或者染色体位点，如果我们要研究某一特定疾病6个相关的基因，那么我们会在这个数据库中得到6条相应的记录。该数据库允许所有用户查看提交记录。

可以通过网址http: //geneticassociationdb.nih.gov/ 访问该数据库。用户可以在线查询某种特定遗传病相关的基因或某个基因相关的疾病的信息，也可以在免费注册后对整个数据库中的数据进行下载。截至目前，数据库中的记录数已经达到了39 930条。

GAD数据库主要包含三部分功能（位于GAD主页左侧）：数据视图部分；数据查询部分；数据资源部分。数据视图部分主要是为用户从疾病角度、基因角度、SNP角度以及基因与环境互作的角度来查询疾病和基因之间的关联关系。数据查询部分提供了简单搜索、高级搜索、批量搜索以及通过基因来查看所有涉及的疾病的种类和已确实被证明基因和疾病相关的记录。数据资源部分包括用户提交疾病基因关联记录，对GAD数据库的意见以及数据下载等。

首先，用户可以通过数据库页面的左侧的相关链接选择不同的角度对数据表进行查询，GAD会根据用户的查询从数据表中选择相应的字段返回结果页面，并且每条记录的第一个字段都有相应的详细的链接通过该链接，用户可以得到数据表中存储的与查询相关的全部

信息。该数据库中存储的基因（多态）与疾病（表型）间的关系有一部分是通过关联分析得到的，因此数据表中不仅包含显著与疾病发生关联的基因的记录。同时也包含了关联关系不显著的记录，数据表中的字段“Association ？ Y/N”表明了具体的关系，该字段有三种取值：Y、N和空，分别表示该记录的相应研究中的基因与疾病显著关联、不显著关联以及未明确是否关联。以疾病角度查询，我们可以得到特定疾病相关的基因Symbol、染色体区段、基因组定位、对应的OMIM ID、基因与疾病是否关联以及关联显著性水平的p值和相应参考文献的信息，另外GAD还给出了该疾病所属的疾病类信息以及与其他数据库的链接；以基因角度查询，我们可以得到基因相关的疾病表型描述、所属疾病类以及关联显著水平和对应参考文献的信息，同时还可以得到该基因在其他一级基因数据库中的ID、名称、定位等基本信息以及与其他数据库的链接；另外，用户还可以从染色体角度出发，或者通过参考文献、环境因素等方面对数据表进行在线查询，当然，我们也可以选择“All”同时从多个角度对数据库中的相关信息进行查询。

其次，用户可以选择“Simple Search”，利用关键字实现对数据库中相关记录的简单查询。在“Simple Search”中，用户只需要提交以空格分隔的关键字，并选出查询内容的种类（Disease，Gene View，CH-SNP-HapMap和Reference）。还可以选择“Advanced Search”增加查询限定条件进行数据记录的高级搜索，包括更新时间，与疾病是否关联，疾病表型，疾病种类等。如果某些限定条件选择空白则会列出相关条件下的所有记录。

GAD还支持对基因的批量查询，用户可以把小于300个基因以HUGO中的基因Symbol，UNIGENE ID，或ENTREZ GENE ID的形式形成一个基因列表，并通过该列表实现对GAD中信息的批量查询。这样，GAD就可以分析高通量实验（microarray、cDNA sequencing、SAGE等）得到的基因与人类疾病之间的关系。

选择“Browser All”链接可以得到结果，返回了数据库中的所有基因和与各类疾病间的关系，如第一条记录HESX1基因，它在数据库中共存在3条相关记录，其中与代谢类疾病（MET）相关的记录有1条，与其他类疾病相关的记录有2条。

用户还可以选择“Positive Only”以筛选得到疾病与基因间存在显著关联的记录。

同时，用户还可以通过“Add Record”页面实现向数据库中提交记录；通过“Download”页面实现对数据库中数据的下载。

目前该数据库已经得到了研究人员的广泛应用，例如，2009年Liu等人发表在BMC Bioinformatics杂志上的文章“The ‘etiome’: identification and clustering of human disease etiological factors”中，作者为了研究影响疾病的因素从GAD数据库中获取了与1034种复杂疾病相关的1100个基因的相关数据；2008年Yang等人发表于BMC Bioinformatics杂志上的文章“An integrated database-pipeline system for studying single nucleotide polymorphisms and diseases”为了得到一个可用于研究遗传变异与疾病间关系，也从GAD中提取了疾病相关信息进行数据整合。

（三）癌症基因组剖析计划数据库（CGAP）

癌基因组剖析计划（cancer genome anatomy project，CGAP），是一项由美国癌症研究所（national cancer institute，NCI）于1996年发起并建立和主持的交叉学科计划。其目的在于产生用于解码肿瘤细胞的分子结构所需的信息，并创建一系列技术工具以挖掘与肿瘤相关的

基因、蛋白及其他的生物标记，最终为癌症的研究提供信息资源和技术方法。CGAP的总体目标是检测正常、癌前病变以及癌细胞的基因表达谱，使得研究人员可以借助于这些表达数据描述出肿瘤形成过程中的一系列细胞分子特征，最终改善对患者的检测、诊断和治疗。该计划通过与全世界范围内科学家的合作来增强其信息的科学性和完整性，为癌症相关科研人员提供方便。

CGAP被分为五个部分，每一个都有它自己的目的、信息学工具和资源。人类肿瘤基因索引（the human tumor gene index，hTGI）指明了在人类肿瘤发生过程中的基因表达；分子表达谱（molecular profiling，MP）从分子水平分析人类组织样本的概念；癌症染色体变异计划（the cancer chromosome aberration project，CCAP）描述了同恶性转移相关的染色体变异；遗传注解索引（the genetic annotation index，GAI）指明和描绘了同癌症相关的多态性；小鼠肿瘤基因索引（the mouse tumor gene index，mTGI）确定了在小鼠肿瘤发生过程中的基因表达。

用户可以通过该网址http: //cgap.nci.nih.gov/对CGAP的网站进行访问，并通过左侧导航栏CGAP Info中的相关链接了解更多有关该计划的更为详细信息。该网站提供了七个相关模块用以对所有CGAP中包含的数据、生物信息学分析工具以及生物学相关资源的查询和获取，借助于这些模块用户可以实现对生物学问题的计算机模拟，从而快速地获得问题的解决方案。进入"Genes"的标签页，可以得到页面，该页面中提供了多种可用于对癌症相关基因进行查询和分析的工具，如利用"Batch Gene Finder"可以实现对多个基因的批量查询，利用"Nucleotide BLAST"工具可以找出给定核苷酸序列中最有可能的候选基因等，对于查询到的每个基因，CGAP都会提供一个包含NCBI以及NCI的多个子库中有关该基因的描述信息在内的"Gene Info"页面。

下面我们以使用Gene Finder工具为例简要介绍如何在CGAP中实现对癌症相关基因的查询，并对查询结果进行简要解释。

Gene Finder对应的标签，用户可以利用该工具通过输入某个特定基因的Gene Symbol、GenBank数据库中的accession number、UniGene数据库中的cluster ID 或者Entrez Gene ID来查询基因的相关信息。也可以通过限定组织、功能、定位等方面的条件来实现对相关基因的查询。例如要查询与人类的结肠（colon）组织相关的基因，我们首先应在选择物种（Select organism）这一下拉列表中选择"Homo sapiens"（目前CGAP只支持对人类和小鼠Mus musculus两个物种的查询），并在Tissue Type对应的下拉列表中选择Colon，提交查询后可返回一个包含所有结果的Gene List页面，对于感兴趣的基因，我们还可以通过页面中对应记录的最后一栏"Gene Info"链接去获取有关该基因的更为详细的信息。对于结果列表中的第一个基因A1CF，CGAP中包含的有关该基因的全部信息，其中包含A1CF在其他数据库中的ID名称，并提供其他数据库对该基因的描述链接，同时还包含了A1CF相关的序列、表达、细胞遗传学定位、染色体定位、对应蛋白、同源物以及相关的GO注释等多方面的信息。

CGAP中还包含有染色体（Chromosomes）、组织（Tissues）、SAGE精灵（SAGE Genie）、通路（Pathways）、工具（Tools）和RNA干扰（RNAi）六个模块。与Genes模块类似，每个模块都提供了很多相关的查询分析工具，可支持对CGAP中包含的染色体畸变、表达数据、蛋白复合物、生物学通路等信息在内的多方面内容进行搜索，并可以根据查询得到的结果做进一步更深入的分析研究。特别是RNAi模块，其中收录了靶向癌症相关基因的RNA干扰结构，并包含有已经证实的靶向癌基因的短发卡RNA（short hairpin RNA，shRNA）。

另外，CGAP还允许用户对其中的数据资源进行下载，在CGAP主页左侧导航栏中包含有CGAP Data项，该项中的内容Download就是数据下载页面的链接，下载页面包含了人和小鼠两个物种的基因注释、基因表达以及相关的一些文库中的数据。

CGAP计划还有另外一个目标，就是建立一套完整的基因及其变异目录，这些目录不仅有利于评价癌症的危险程度，而且可以根据遗传变异确定预防或治疗策略，最终根据分子特征达到治疗的目的。目前CGAP建立的注释基因索引包括利用表达序列标签（Expressed Sequence Tags，EST）及基因注释等途径建立的人和小鼠的肿瘤基因索引和用于区分鉴定与肿瘤有关的基因的遗传变异的注释索引。CGAP还建立了许多cDNA文库，不仅包括有全瘤组织文库，也包括癌症发展过程中不同阶段的细胞cDNA文库。同时CGAP也提供了诸多资源如克隆、BAC及技术方法和检索工具等，为肿瘤研究提供了一个多学科的综合平台。

CGAP蕴涵了大量有用的信息，目前，已有许多科研工作者成功地利用这些信息实现了对肿瘤的研究，如Loging等用数据库及快速表达筛选方法，通过CGAP鉴定胶质瘤潜在的肿瘤标志和肿瘤抗原，获得了有意义的结果。

第二节

变异组信息学与复杂疾病遗传定位

Section 2 Genome Variations and Complex Disease Mapping

一、变异组学与人类疾病

人类疾病的发生是多种因素共同作用的结果。绝大多数常见疾病，如糖尿病、癌症、心脏病、精神性疾病等具有非常强的家族聚集特征，表明遗传因素在疾病形成中有重要作用；而同一家族某些成员发病、另一些成员不发病，以及同一种疾病在不同个体中具有不同的严重程度和表现症状，这些又体现了常见疾病的多因素特征。事实上，现有的研究提供了大量的证据显示常见疾病遗传上的复杂性，认为常见疾病是众多基因共同作用的结果，而且人与人之间在疾病发生中的差异很大程度上可以通过遗传变异来解释，并在此基础上提出著名的“常见疾病，常见变异”假说。

我们知道，任意两个不相关个体的DNA序列有99.8%是一致的，而剩下的0.2%由于包含了遗传上的差异因素，造成人们不同的生理表型、罹患疾病的风险及不同的药物反应，这些差异在人类多样性形成中也具有同等重要的意义。这0.2%的差异在基因组序列中具有不同的类型和作用形式。其中，不同个体DNA序列上的单个碱基的差异，称作单核苷酸多态性（single nucleotide polymorphisms，SNPs，图9-2A），例如，某些人的染色体上某个位置的碱基是A，而另一些人的染色体的相同位置上的碱基则是G，而同一位置上的每个碱基类型叫做一个等位（allele），除性染色体外，每个人体内的染色体都有两份，即我们常说的同源染色体，一对同源染色体上的两个等位的组合叫做基因型（genotype，图9-2B）。对上述SNP位点而言，一个人的基因型有三种可能性，分别是AA，AG或GG。而检定基因型的过程，称作基因分型（genotyping）。由于SNP在人群中具有最大的数量和最广泛的分布，且易于分型，已经成为现代遗传变异与复杂性状研究中最重要的研究对象，也是生物医学、农业、畜牧业研究中非常重要的研究工具。

如果将世界上所有人看作一个群体，那么全人类中大约存在一千万个SNP位点，这些SNP绝大多数呈现二态性，并且具有不同的等位频率，我们将在某个研究群体中出现较少的等位频率称作最小等位频率（minor allele frequency，MAF），并以此将SNP划分为常见和罕见两类，一般说来，常见的SNP最小等位频率应当大于5%（也有文献定为1%），具有比较广泛的群体分布，与个体表型差异和疾病易感有关；而罕见的SNP往往是某些单基因病或偶发疾病的承载者。由于减数分裂过程中，染色体发生重组的位置具有选择性，染色体上距离越近的SNP越倾向于以一个整体遗传给后代，这样，我们把位于染色体上某一区域的一组相互关联

的SNP称作一个连锁块(linkage block),这是我们将SNP作为一种重要的遗传标记进行复杂性状和复杂疾病定位的分子基础。

除了从频率的角度对SNP进行划分,并在此基础上进行基于统计思想的遗传定位分析

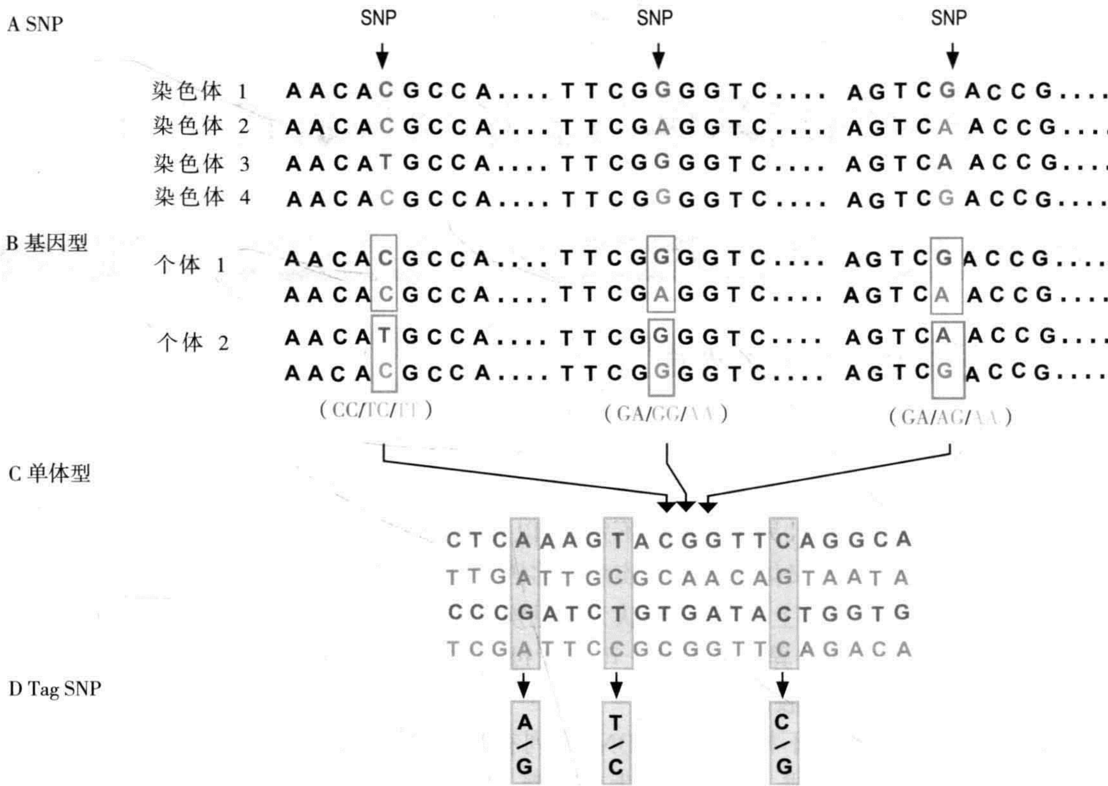


图9-2 SNP、基因型、单体型与Tag SNP

A图中彩色标记出不同的SNP位点,及其在不同个体中的等位情况;B图显示同一个体某个基因座上两个等位位点组合,即基因型;C图中将某个个体的同一条染色体上的SNP放在一起,将其定义为单体型,这里的单体型是一个狭义的概念,也是本章研究的单体型含义;D图是在单体型基础上提出的基于群体分布的单体型标签,即Tag SNP

外,由于SNP本身数量众多、分布广泛等特点,它还具有非常重要的功能特性。我们习惯于将分布在基因(编码或非编码)区域,并且能够直接影响基因表达数量或基因产物(蛋白质或RNA)结构的SNP称为非同义SNP(non-synonymous SNP)。在实际研究中,还发现不同SNP之间具有潜在的相互联系,同一个基因或同一个生物学过程中多个SNP的互相作用能够起到从量变到质变的效果,直接影响生理指标、病理发生和药物反应的差异性。这些提示我们从功能和生物学系统的角度研究SNP在复杂性状和复杂疾病中的作用非常重要。

人类的遗传变异是多样的,有些变异之间也许可以通过连锁不平衡原理由SNP进行发现和解释,但有些变异本身行使着复杂的生理和病理学功能,是SNP所不能替代的。这里简要介绍一下人类染色体中其他的遗传变异,涉及最简单的变异形式插入/删除多态(In/Del)、关系碱基数量最大的多态拷贝数变异(copy number variants, CNV)、早期应用的遗传标记微卫星(microsatellite, MS)等。

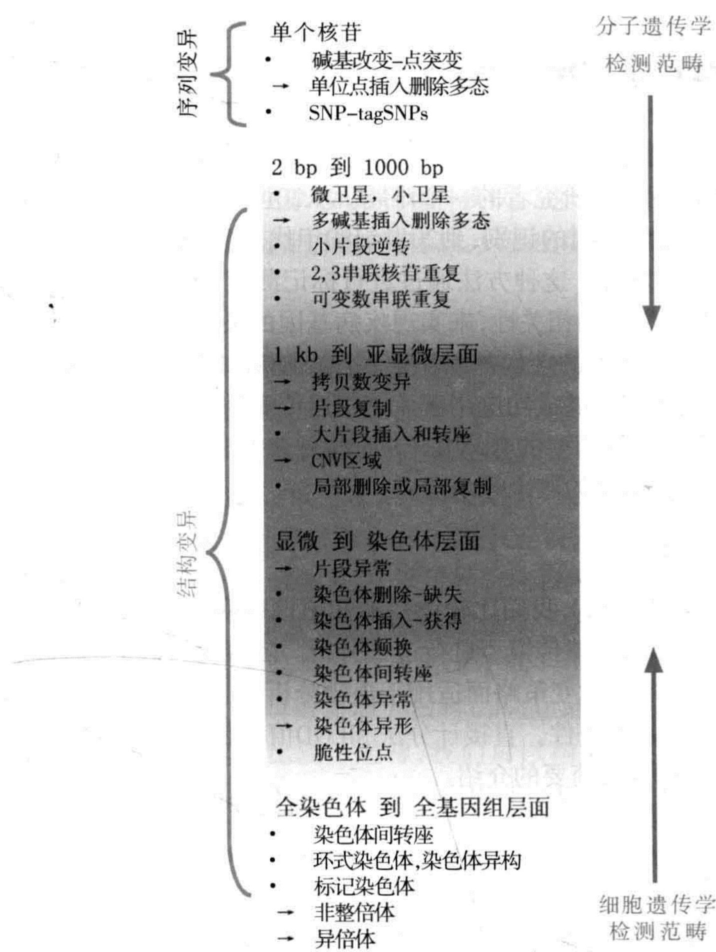


图9-3　人类染色体上的序列和结构变异

如图9-3所描述的人类染色体上的各种遗传变异，我们以1kb长度为界，将遗传变异分为两类，一类自身影响的范围比较小，是包括SNP在内的序列变异，另一类是从微卫星和插入删除多态起到长重复片段的结构变异，更大的染色体变化我们将之为染色体畸变，也是遗传学研究中的重要范畴，这里我们不展开介绍。

微卫星多态目前已发现5000余个，是早期遗传定位研究中非常重要的分子标记，也与癌症等多种疾病的稳定性有关。已经发现的人类插入删除多态已达到586个，这些多态最长能达到70kb，在多种疾病，特别是精神病发生过程中有重要的作用。CNV目前已经识别了1447个，涉及360Mb的染色体范围，占人类色体总量的12%，是影响核苷酸数最多的变异形式。由于CNV本身的长度超过100kb，能够直接引起基因拷贝数、调控区段的变化，因此对于生理病理有着重要的影响。变异组学的研究证据不断的告诉我们，人类染色体中还有着巨大的未知的秘密，既决定了人类种族的一致性，又决定着人类多样性的产生，由于他们的存在，这个世界变得绚烂多彩，同样由于他们存在，人们对人生的感悟又有所不同。真正全面了解这些遗传变异在人类生理病理中发挥的重要作用，才能够实现从系统的角度揭示人类生命的本质。

二、SNP与人类复杂疾病定位

复杂疾病机制研究是生物医学研究中的重中之重。致病基因的发现是研究复杂疾病机制的重要环节，也是长期困扰科学研究者的一个难题。从20世纪初，人们就在探索基于分子标记的统计分析方法用于致病基因的识别，到20世纪80年代，伴随分子生物学技术的革新，这一研究方案得到了长足的发展。这种方法通过进行标记测定，采用统计学方法研究分子标记的遗传特性与疾病发生之间的相关性，来实现疾病基因的染色体定位，而几乎不需要任何先验的生物学知识，是一种强大的疾病基因识别手段。随着SNP分型技术的发展，SNP作为一种最重要的分子标记，不仅能够成功应用于孟德尔遗传病的研究，同时被广泛用来进行复杂疾病的染色体定位。本节将简要的介绍基于SNP的复杂疾病遗传定位实验样本选取准则、连锁分析、关联分析、统计结果的取舍等内容。

(一) 参数连锁分析方法

对于孟德尔遗传病(单基因病)，我们比较清楚地知道该疾病的遗传方式、外显率、基因频率等指标，从而确定一个准确的遗传模型进行连锁分析。随着统计方法的不断发展，某些遗传模型并不清楚的疾病也通过改变策略而适用于连锁分析，但无论如何，相对准确的模型建立是参数连锁分析成功的基本条件。直接计分法和LOD值法是最常用的参数连锁定位方法，这里我们以LOD值法为例进行简要的介绍。

LOD值法进行连锁分析首先针对某一疾病收集一定数量的家系资料并进行分离分析，确定遗传模型；然后通过文献检索了解其可能的决定性状的染色体区域，并对该区域的SNP进行查询和筛选，基于选定的SNP，对该家系成员进行基因分型；最后通过连锁分析估计疾病与SNP在子代中重组的发生率，计算LOD值，确定重组分数及相应的遗传距离，并进行假设检验，判断易感基因是否与遗传标记连锁。

LOD值是指在一定重组率θ条件下，两个位点相连锁的似然性和不连锁的似然性比值的对数值。即：

$$LOD=\log_{10}\frac{\text{两位点连锁的似然性}}{\text{两位点不连锁的似然性}} \tag{9-1}$$

在进行连锁分析时，要计算$\theta=0.0$(不重组)到$\theta=0.5$(随机分配)的一系列LOD得分。当LOD得分为+3或更大时，肯定连锁；当LOD值得分小于或等于-2时，排除连锁。LOD值得分最大时的θ值被接受为最大似然估计值。由于现有的LIPED(http://linkage.rockefeller.edu/ott/liped.html)、LINKAGE(http://linkage.rockefeller.edu/soft/linkage/)、S.A.G.E.(http://darwin.cwru.edu/sage/)等自由软件包提供了包括LOD值法在内的多种参数连锁分析工具，这里对具体的算法不再展开。由于早期的连锁分析方法对模型的依赖性较强，主要适用于单基因病，计算速度慢等原因，新的方法也在不断的开发，如“混合模型”方法、多位点连锁分析方法、基于仿真的吉布斯取样及蒙特卡罗方法等。

参数连锁分析方法已经被应用于几百种孟德尔遗传病的遗传定位研究中，同时也在某些复杂疾病研究，特别是大家系研究中获得成功。当然，实际的疾病家系非常复杂，所以在研究中还应该注意一些特殊的情况：①如果在特定的家系中难以获得明确的连锁关系，还可以收集大量的家系资料进行分析，但并不是说连锁分析结果在某些家系中出现阳性结果就可以忽略

阴性结果的家系，背后可能还存在更复杂的遗传机制。同样，在实验样本获取部分我们曾经提出五个基本的原则，参数连锁分析家系选择过程中也可以考虑以上的因素，做出合理的家系筛选。②对于某些外显率并不明确的疾病，还需要对外显率进行估计，而采用疾病个体特异的分析策略，将无病个体设置为表型未知个体也是一种有效的分析方法。③家系中某些个体的疾病表型并不典型，难以确定是否受累，如某些精神疾病。这时就需要进一步严格疾病定义，将出现某一特定的表型作为诊断的标准，或放宽标准，只要出现疾病某一典型表现即定义为受累。

(二) 非参数连锁分析方法

非参数连锁分析是一种在分析前不需要确定疾病遗传模式（如基因型频率、外显率等）或半依赖模型的分析方法。最常用的非参数连锁分析方法是等位共享方法。等位共享方法不依赖于遗传模型的构建，而是一个排除模型的过程。通过显示受累亲属间高于随机情况的共享遗传相同的染色体区域（或位点）概率来证实染色体区域的遗传模式与孟德尔遗传之间的差别。由于等位共享的方法是一种非参数方法，比参数连锁分析方法有更宽泛的应用范围，而且即使在受累亲属中不完全显性、表型复制、遗传异质性和高频等位等影响因素存在时，也有较好的表现。而唯一的缺陷是等位共享方法提供的结果一般说来没有参数连锁分析方法显著。

等位共享方法研究家系中亲属在共享来源于同一祖先的特定染色体区域或位点的频率，我们把这种区域或位点也叫做血源一致性（identical-by-descent，IBD），然后将某个位点共享IBD的情况与随机进行比较。通常，我们可以构建一个血缘一致性受累家系成员（identity-by-descent affected-pedigree-member，IBD-APM）统计量：

$$t(s)=\sum_{i,j}X_{ij}(s) \tag{9-2}$$

式9-2中，$X_{ij}(s)$是指家系中第i个和第j个亲属在染色体位点s处共享IBD的个数，加和指的是这个家系中所有亲属对在s处共享IBD的个数。如果是多个家系的组合研究，那么可以加和成$T(s)$。在随机分离状态下，$T(s)$趋于均值为μ，标准差为σ的正态分布，μ和σ可以通过计算血缘系数（kinship coefficient）获得。当统计量$(T-\mu)/\sigma$超出了设定的阈值，我们就可以判定此时的状态与随机分离相偏离，从而得到阳性的结果。

在等位共享分析中，最简单的一种形式是同胞对（sib pairs）分析，同胞对共享IBD数为0，1或2（随机情况下，共享频率分别为25%、50%、25%，图9-4），可以采用简单的x^2检验分析疾病状态下的等位共享情况。这样的方法同样可用于受累叔侄对、表兄弟对的研究。

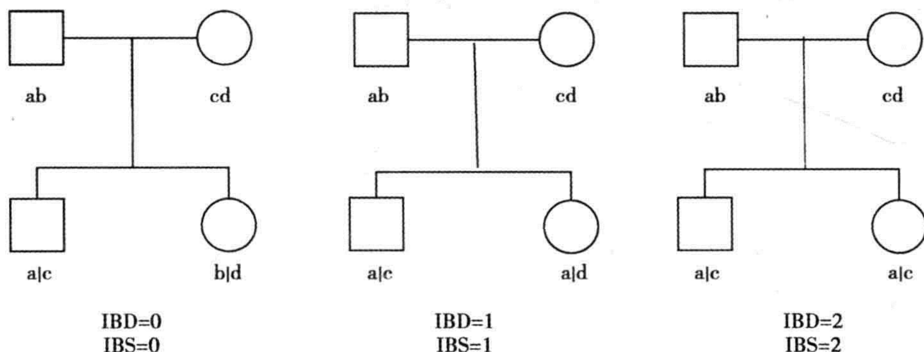


图9-4 同胞对血缘一致性和状态一致性示意图

IBD之外，还有一个与之相似的概念状态一致性（identical-by-state，IBS）。IBS用来描述亲属对之间共享同一等位（不区分是否同一祖先来源）的频率。两者的基本分析方法是相通的，但采用IBS方法可以避免IBD-PAM分析过程中对IBD的估计过程，因此应用也非常广泛。随着遗传标记分型技术，特别是SNP分型技术的进步，IBD和IBS方法也逐渐应用于基因组范围关联研究中。

（三）关联研究发现疾病风险SNP

关联研究（association study）是不依赖于家系信息的一种遗传定位策略，由于资源丰富，分析方法简便，是目前遗传定位研究中最常用的分析方法。关联研究通过检验某个特定的等位在疾病组和对照组中出现的频率差异来判断此等位是否是疾病易感等位。以SNP而言，发现风险SNP的过程可以采用四格表x^2检验进行等位频率分析，也可以采用$2 \times 3x^2$检验进行基因型分析。

某医院对200名高血压患者和200名对照个体进行检测，通过限制性内切酶方法对采自这些个体的外周血淋巴细胞进行分析，获得了SNP rs39461的基因型（表9-2），假定此次研究不存在采样上的缺陷，问这个SNP是否与高血压的发生相关？

表9-2 患者及对照个体的基因型统计表

分组	基因型			合计
	CC	CT	TT	
疾病组	3	36	161	200
对照组	3	57	140	200
合计	6	93	301	400

在一般的SNP分型实验中，我们首先获得的数据就是个体的基因型数据，对这些个体按疾病和对照组进行统计就能得到类似于表9-2的统计表格。根据学过的统计学知识，我们知道，这个例题事实是一个两样本频数（计数资料）差异比较问题，如果直接从基因型频率考虑，这个问题适用于自由度为2的卡方检验，那么，我们可以进行这样的处理：

（1）建立检验假设，确定检验水准

H_0：在检测群体中，这个SNP与高血压的发生相关

H_1：在检测群体中，这个SNP与高血压的发生不相关

$\alpha=0.05$

（2）计算检验统计量

$$\chi^2 = n\left(\sum\frac{A^2}{n_R n_C}-1\right), v=(R-1)(C-1) \quad (9-3)$$

n为总例数，R、C分别为行数和列数，A^2为每格的频数，$x^2=0.45$为自由度，将表格中各数值代入公式得$x^2=0.45$，$v=2$。

（3）确定p值，作出推论

查表得$p=0.746>0.05$，按$\alpha=0.05$的水准，接受H_0，即在此检测群体中，SNP rs39461与高

血压的发生没有相关性。

在上面的例题中，我们采用了简单的统计方法对SNP与疾病关联性进行了分析，方法上的简捷性显而易见。但关联研究也有比较明显的缺点，即对对照组样本选取具有严格的限制，此外，由于关联研究可能针对任何一个分子标记进行，而不存在先验的假设，对关联研究发现的风险SNP尚需要进行可靠的功能验证。由此可见，关联研究中对标记信息的分析比研究方法本身更重要，下面我们将从关联研究机制上来探讨风险SNP发现应注意的问题。

关联研究中发现SNP与疾病发生之间的显著相关性可能存在三个原因：①SNP本身就是一个致病的SNP；②SNP本身不能导致疾病，但与导致疾病的基因处于连锁不平衡状态；③研究群体选择失误造成的统计显著性。第三种情况是关联研究过程中需要避免的，所以关联研究过程中还应注意三点：①关联分析的样本选取要严格限制在同质性群体中；②关联研究对照组选取应当谨慎，必要时选择未受累亲属作为内对照；③如条件允许，对于获得的阳性位点可进行传递不平衡检验（transmission disequilibrium test，TDT）来确认发现的致病等位在家庭遗传中倾向于向患病子代遗传。

由于复杂疾病发生过程中，存在遗传位点间的相互作用，单个位点的关联分析方法有时不能获得足够的信息来发现某些区域与疾病之间的关联性。基于单体型、罗杰斯特回归、主成分分析、随机森林等统计学和机器学习方法的遗传定位方法成为有用的研究手段，得到了比较广泛的应用。

总起来看，关联研究和连锁分析有很多重要的区别。关联研究检验疾病与等位频率在群体中是否存在相关性，连锁分析检验疾病与位点是否在家系中共同传递。当群体中致病因素是多样的，而且致病位点相互独立，散在存在的时候，每个位点与疾病关联都将很弱，遗传定位中往往只能检测到连锁而难以发现关联；相反，当致病位点等位效应较弱，对疾病贡献较小时，但在疾病个体中有较高的等位频率时，基于家系的连锁分析难以发现潜在的传递模式，而关联研究却能识别出这种致病位点。因此，关联研究和连锁研究本身并不存在孰强孰弱，而需要考虑实际解决的问题进行选择。

传统的连锁和关联分析依赖于实验室SNP分型技术，如限制性片段长度多态性方法、变性梯度凝胶电泳、等位基因特异寡核苷酸片段分析等，伴随高密度基因芯片技术的发展，这些技术对于测序低通量或单基因多态位点有着各自的优势，经济实用，便于一般实验室从头设计基于单基因或某一染色体候选区段的风险SNP筛选。伴随新型高密度基因芯片技术的发展和商业化，单次实验可对某个样本数十万甚至上百万的SNP位点进行同时测定。单次实验SNP测定数量的增加，使得人们有可能从更大范围，直至全基因组范围进行疾病关联的SNP筛查，并将关联分析或连锁分析方法扩展到整个基因组维度，即目前广泛开展的基因组范围关联分析（genome-wide association study，GWAS）。

（四）遗传分析中的统计显著性

遗传分析方法虽然笼统的分为两类，但相应的研究方法众多，既有传统的统计分析方法，也有衍生而来的机器学习方法，但无论采用何种方法进行复杂疾病的遗传分析，最终都将面对统计结果的取舍问题，即如何进行统计显著性的阈值设定。而且，这个问题，还将因为遗传分析中分子标记的增多或检验模型的增加，特别是GWAS的开展而变得更

为严峻。

在进行SNP与疾病之间的连锁或关联分析时，我们要设置一个可以接受的假设检验显著性水平α（一般为5%）。这样，每一次检验，都有5%的可能引入一个假阳性的结果（Ⅰ类错误）。当进行n次独立的连锁或关联检验时，引入的I类错误水平将满足$\alpha=1-(1-\alpha')^n$，当n变大时，引入的假阳性结果也将增多，从而使得在进行数以千计的SNP关联或连锁分析时，需要对α进行Bonferroni校正$\alpha'=\alpha/n$。在这种情况下，如果对1000个SNP进行检验，且要达到显著性水平$\alpha=0.05$，需要达到真实的显著性水平为$\alpha=5\times10^{-5}$，而100万个SNP进行检验时，所需要达到的真实显著性水平为$\alpha'=5\times10^{-8}$，这对于高维度SNP遗传定位是个灾难性的结果，直接导致单次关联或连锁分析所能获得的显著性结果极少，一方面许多真正相关的SNP没有被发现，造成了很大的假阴性，另一方面在发现的极少的显著性结果中依然存在着较大的假阳性。

因此，对于遗传定位的结果取舍，特别是多重检验问题一向都是人们关注的重点，采用多次随机进行SNP与疾病相关性检验进行显著性水平选取是目前为回避多重检验校正而广泛采用的一种方法。另外，考虑到基因组中广泛存在的连锁不平衡问题，对待检的SNP进行LD修正是降低多重检验校正影响的一种有效方法。此外，在芯片分析中采用的FDR方法也经常用于遗传定位结果的修正。

三、变异组学研究资源

（一）国际人类单体型图计划及其应用

1. 国际人类单体型图计划概况　国际人类单体型图计划（international HapMap project, HapMap）是继国际人类基因组计划之后，人类基因组研究领域的又一个重大国际合作项目。HapMap计划起始于2002年，由美、加、中、日、英、尼日利亚等国研究机构发起、参与及完成。中国科学家承担3号、21号和8号染色体短臂单体型图的构建，工作量约占总计划的10%。项目共取样270个正常个体，其中有欧裔美国人和尼日利亚雅鲁巴人（非洲）各30个核心家系（90个个体），及中国北京汉族人及日本东京人各45个个体。一期已于2005年完成，成功分型100多万个常见SNP位点的识别，达到平均每3kb一个SNP的测定。由于染色体连锁不平衡的存在，一期数据可以捕获基因组上80%的遗传差异信息。二期计划在一期基础上完成300多万个SNP位点的分型，构建起一张精度更高、信息更完整的多人种遗传多态图谱。三期计划已经开展，在进一步测定原有群体基因型基础上，加入另外7个不同历史遗传背景的人群，部分分型数据已经发布。HapMap计划期望在全部完成时能够提供一个包括全部人类遗传差异的多态组图谱，同时带动其他人类遗传变异的发现和研究。

2. HapMap数据特点与扩展应用　HapMap计划建立了人类全基因组遗传多态图谱，依据这张图谱我们可以进一步研究基因组的结构特点以及SNP位点在人群间的分布情况，为群体遗传学、进化遗传学分析提供数据，也为复杂疾病的遗传定位提供高密度的SNP数据参考。HapMap的构建分为三个步骤：①在多个个体的DNA样品中鉴定单核苷酸多态（SNP）；

②将群体中频率大于1%的那些共同遗传的相邻SNP组合成单体型；③在单体型中找出用于识别这些单体型的标签SNP。这样，HapMap提供的每个研究个体的数据包括SNP等位、基因型、基因型频率、200kb范围内SNP之间的LD量度（r^2、D'）。

伴随HapMap计划的进一步拓展，结合群体遗传学的研究手段，我们可以更加深入地去观察和研究基因组。基于大群体、多种群的人类单核苷酸多态数据的重组率推算提供了我们一张基因组进化痕迹图；连锁不平衡的计算给了我们一张基因组块状连锁结构图；种群差异研究让我们看到一张种群间基因组结构差异图；SNP的杂合情况告诉我们人类基因组上受到选择的区域或区域内的基因；利用SNP位点向两边延伸的长度差异情况，我们可以观察到一些基因组上近期正在进行的选择事件，甚至是当前正在悄悄进行中的进化，因为新产生的突变位点传代较少，它和周围位点的连锁情况受重组事件的影响较小，另一方面优势突变也会因选择压力的存在使周围的重组受到影响……当然这些不同的指标中也隐藏了人类成长过程中的一些信息，例如迁徙、战争、灾难、繁盛等对基因组遗传多态性产生影响的历史事件。

此外，高密度的SNP位点，为进一步加强和完善基因组范围的表型和遗传相关性分析（关联研究或数量性状定位）提供了可能，以往遗传学上定位基因使用较多的工具是微卫星，这些新产生的SNP位点弥补了微卫星在基因组上分布不够均匀、密度不够高的缺点，是一种更为有效的分子标记。目前，已经有很多致病基因借助SNP数据得到定位。另外，根据SNP在基因的不同功能元件中的分布情况和基因在细胞中的表达情况，我们可以研究基因上的不同元件序列是如何控制蛋白表达进而影响个体表型的。伴随着HapMap三期数据的产出、各种实验技术的进一步发展，以及更加大量的基因组序列数据加入到人类的知识库中，与此相关的研究方法和研究手段会不断出现，我们将能够更加完整、更加深入、正确地认识我们自己，揭示生老病死的奥秘，并为人类生存质量的提高提供有益的参考信息。

3. 利用HapMart进行科学研究　为了便于科研工作者快速提取感兴趣的SNP数据，在HapMap数据基础上，BioMart（一个重要的生物信息学数据分析平台）开发了方便、友好的SNP获取网络平台HapMart。这个平台支持研究者输入SNP、基因、染色体区段等信息进行限定条件下的SNP查询及相关信息的输出。由于HapMap数据本身跨群体的特性，用户可以通过这个平台进行不同群体间的数据提取，如果是候选基因或多SNP实验设计，还可以联系其他的连锁不平衡分析工具（如下文将提及的Haploview）及感兴趣的基因型频率信息进行深层次的SNP选择。利用HapMart进行SNP数据的提取主要分为三个步骤：输入设置、输出设置和结果导出。

（1）HapMart的输入设置：图9-5显示了HapMart查询过程中的输入和查询限制界面。在这里，可以进行研究群体的选择、SNP质量限定，以及查询设置。目前，HapMart主要支持四个群体的查询，后续的群体正在添加中。对于目标SNP可以进行最小等位频率、分型机构、分型平台、SNP类型的限定。可以根据SNP的标识符、定位区域（功能区域或染色体位置），及其与基因的位置关系进行单个或高通量的SNP查询。

（2）HapMart的输出设置：图9-6显示了HapMart的SNP输出属性设置界面。可以根据研究者的研究兴趣进行设定，并输出相应的结果。SNP相关属性主要有标识、遗传定位、等位和基因型状态和频率特征。

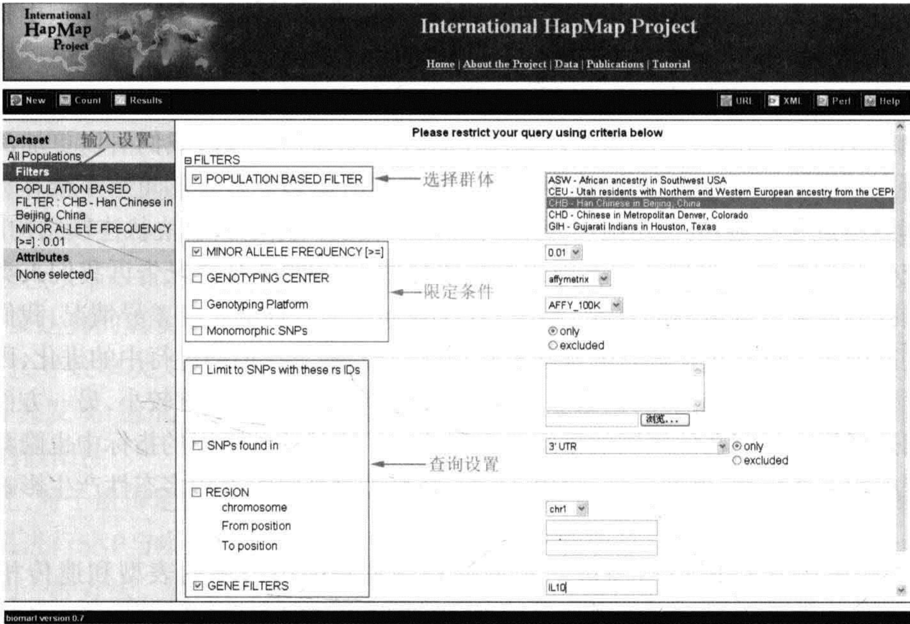


图9-5　HapMart的输入设置界面



图9-6　HapMart的输出设置界面

（3）查询结果的导出：根据研究者的研究兴趣和输入输出设置，以特定的格式显示和导出查询结果。图9-7显示限定最小等位频率0.01时，定位在基因IL10上的SNP位置、等位和基因型频率信息。

HapMart查询结果以HapMap数据为基础，提供的是不同种群特定群体的SNP信息，主要用以实验设计者针对特定人群的实验参考。由于计划测定规模的限制，数据本身存在一定的偏差，因此查询结果应当进行一定的预实验和初步分析，才能用于大规模实验。



图9-7 HapMart的结果显示与导出界面

(二) SNP存储与维护数据库dbSNP

SNP作为新一代遗传标记具有数量多、分布广、密度大等特点，已广泛应用于遗传学研究中。为了满足对基因组范围总体变异的需求，解决在关联研究、基因定位、功能和药理遗传学、群体遗传学、进化生物学以及定位克隆、物理作图等领域中大规模抽样设计的需求，NCBI与NHGRI协作创建了dbSNP。通过dbSNP，由公共和私人组织提交的遗传变异数据与其他信息来源，如GeneBank、PubMed、LocusLink及人类基因组数据实现交叉引用，为广大研究者提供了丰富的遗传变异，特别是SNP信息，呈现了一幅全面的人类SNP的基因组分布图。充分利用数据库中资源将大幅度降低研究成本、提高研究效率。此处，就dbSNP数据库的功能、范围、数据提交、检索进行简要的介绍。

1. dbSNP的主要功能

（1）遗传变异序列环境分析：dbSNP通过BLAST和E-PCR对变异周围序列进行分析，将其链接到其他NCBI序列资源，对变异进行交叉注释。用户可直接在dbSNP中检索，或在NCBI查询空间的任何部分开始，构建一个满足要求的dbSNP记录集，该记录可通过超文本或URL与外部信息资源整合。

（2）基于NCBI的遗传变异交叉注释：在后基因组时代，对特征序列的注释（如新基因或调控区域）为当前在随机序列中发现的变异提供一个功能背景。随着这些新基因条目的出现，dbSNP通过链接能够将变异自动注释到恰当的参考序列集或UniGene集中。

（3）外部资源整合：dbSNP具有“LinkOut URLs”功能，将变异信息链接到NCBI之外的信息资源。这种整合非常重要，尤其是当我们考虑将变异注释到整个基因组上或考虑其对生物体的意义时。

（4）遗传变异的功能分析：NCBI没有直接地在序列上注释变异的详细生物化学或者表型信息，而在dbSNP中保留了与外部数据库的链接。因此，dbSNP记录能够链接到那些对个别变异描述更加完整的位点特异突变数据库。

2. dbSNP数据特征 dbSNP数据库中不仅收录了人类SNP数据，还收录了所有已知的跨物种的SNP、插入/缺失、拷贝数和微卫星多态，且包含种族特异频率和基因型数据、实验条件、分子背景，以及功能特性和临床变异的定位信息。截止到2009年10月7日，dbSNP已经更新至130版本，涉及55个物种的1.5亿个SNP，编码区SNP已超过2千万，具有频率信息的SNP超过300万个。

3. 向dbSNP提交数据 目前，科研领域出版物中涉及的遗传变异信息一般要求提交到dbSNP数据库中。所需数据提交信息包括特定位点观察到的等位基因、突变周围的侧翼序列、使用的实验方法，伴有STS或GeneBank记录的指针。每个特异实验室具有唯一标识，这将允许提交的数据与特定试验室相关联。NCBI将会给每个提交的SNP分配一个编号ss#，一种生物基因组中涉及的唯一SNP也将分配一个标识符（人类的SNP标识符为rs#）。所有这些编号或标识符被用于将SNP映射到外部资源或数据库中，包括NCBI中其他数据库。

4. 利用dbSNP进行信息检索 在dbSNP中可直接查询，也可通过其他NCBI查询框来检索。直接查询可以通过提交实验室、新的批量提交、鉴定方法、群体类型研究、书刊题目、群体变异水平或STS映射信息实现。作为NCBI中一个整合部分，dbSNP中内容与其他信息资源记录是横向链接的。从其中任何来源中查询的结果集合会给用户提供一个返回dbSNP相关记录的指针。图9-8显示的是以人类IL10基因相关SNP为例的dbSNP查询过程，及其显示结果，进一步点击蓝色链接将显示每个SNP的详细信息。



图9-8 dbSNP的查询界面

5. 提供dbSNP交叉引用的模块 BLAST: dbSNP查询，可通过标准的BLAST算法来实现，即将用户提交的序列与dbSNP中所有侧翼序列记录进行匹配。除了在NCBI首页中提供了一般的BLAST功能，dbSNP中也提供了此功能。LocusLink: dbSNP也可通过将其与其他NCBI资源整合来检索。通过LocusLink，由基因名字或系统命名来进行检索。从LocusLink数据库中检索的结果将呈现为一个紫色的V形按钮，该按钮可以指向一个LocusLink数据库中任何一个基因上的参考SNP记录列表。Entrez:“图形可视化”旁边的工具条有一个链接将dbSNP中的SNP记录链接到Entrez Gene数据库，这样的链接可以直接看到Entrez Gene中的基因上的

SNP分布情况，并能够根据需求，如是否具有频率、是否编码等信息进行可视化的SNP查询。Genome sequence：重叠视图除了可以设置为显示STS 'marker' 和序列组成，还可以显示"变异体"。

图9-9显示了dbSNP与Entrez Gene之间的交叉引用结果，通过Entrez Gene向dbSNP的超链接，查询到*IL10*基因上的SNP分布情况。dbSNP用不同的颜色和柱体长度表示基因上的SNP类型及其频率状况，对于深入的选取对研究有影响的SNP提供了直观的借鉴信息。从查询结果上看，从dbSNP展开的查询获得的SNP数量要比HapMart的多，这主要是dbSNP本身不限制收录的SNP最小等位频率造成的，所以研究过程中还应当进一步考虑频率和相应的群体信息。

图9-9 dbSNP与Entrez Gene交叉引用

(三)关联研究基因型数据的存储与整理dbGap

1. dbGap的主要功能 国家生物技术信息中心(NCBI)建立了基因型和表型数据库dbGaP。dbGaP的开发是为了存储和发布基因型和表型相关的研究数据及研究结果。这些研究包括全基因组关联研究、医疗测序、分子诊断化验，以及基因型与非临床性状之间的关联性。用于基因分型的高通量、低成本、高效率的分析方法研究，发现海量基因型和表型数据相关性的未知信息提供了强有力的工具。

dbGaP是一个存储了个体水平的表型、基因型和序列数据，以及它们之间的关联性的公共知识库。dbGap收录的数据绝大部分是大规模的基因组范围关联研究数据，对研究过程中得到的信息子集，包括文件、个体表型变量、特征数据表、基因型数据，计算表型与基因型之间的相关性，设定唯一的标识符。其中的部分数据采用直接开放的管理办法，非注册用户即可直接下载相关的数据进行非商业化的科学研究。为了确保被研究者的个人权益及数据检测部门的优先使用权利，dbGap中的大部分数据的访问和使用需要进行人工申请。

dbGaP包含了对各种各样设计的研究。它包括四个基本类型的数据：①研究文件，包括研究说明，协议文件和数据收集文书，如问卷调查表；②每个被评估变量的表型数据，包括在

个体水平上的和以摘要形式进行评估的；③遗传数据，包括研究对象的个体基因型、谱系信息、精细定位结果和重新测序的描述；④统计结果，包括关联和连锁分析结果。

2. dbGap中保存的数据访问　为了保护研究对象的权益，dbGaP只接受已被查明的数据并要求调查员需通过一个授权程序才可以获取个体水平的表型和基因型数据集。总结性的表型和基因型数据，以及研究文件，可以无限制的获取。

dbGaP提供两个访问级别-开放的和受控的-这么做的目的是为了让非敏感数据广泛开放，同时提供对涉及了个人健康信息的敏感数据集进行负责任地监督和调查。研究的总结和测量变量的内容，以及原始研究文件的文本，一般会提供给公众，而要获得这些个体水平的数据，包括表型数据表和基因型数据就需要不同的授权级别。

（1）开放数据：开放式访问数据可以在线浏览或未经批准或授权就可以从dbGaP中下载。这些数据将包括，但并不仅限于表9-3所列的内容。

表9-3　dbGaP中的数据类型

dbGaP 数据类型	信息所在位置
研究	当浏览研究时在名为′ Study′ 的列中出现
	在标签′ Studies′ 下一个搜索的结果
	通往一个变量或一个文件的路径的一部分
研究文件	从′ Browse Studies′ 链接
	与′ Associated Documents′ 下的研究报告链接
	标签′ Study Documents′ 下的一个搜索结果
表型变量	与′ Browse Studies′ 链接
	与′ Associated Variables′ 下的研究报告链接
	标签′ Variables′ 下的一个搜索结果
基因型-表型分析	与′ Associated Analyses′ 下的变量报告链接
	与′ Associated Analyses′ 下的研究报告链接

这是一个可用于开放式进入用户的一般性描述。提供给开放式进入用户的数据可能在研究之间变化，也可能没有通知就与这里描述的有所不同。

（2）受限数据：受控访问数据只能在用户已通过适当的数据访问委员会（DAC）的授权后才能获得。提供给授权的调查人员的数据可能要包括以下内容：①用于个人研究课题的确定的表型和基因型；②谱系；③基因型与表型之间在计算前期的单变量的相关性（如果没有在公开网站上提供）。

由于数据访问策略是基于每个研究的基础上确定的，提供给用户的带有受控访问授权的数据在不同的研究之间可能会发生变化，且也有可能在没有通知的情况下就与这里所描述的有所不同。关于用于一个特定的研究的数据的访问策略，可以在研究报告页连同适当的授权机构的链接上找到更多的细节。

第三节

复杂疾病的分子系统分析

Section 3 The Molecular System Analysis of Complex Diseases

一、面向通路的基因组范围关联研究

随着人类基因组计划和HapMap计划的开展和完成，已识别的人类SNP已达到千万，常见SNP数量也已经达到300万以上，同时HapMap计划推动的商业分型芯片发展，已经促使遗传定位研究由最初的几个至数千个分子标记的研究发展到当前50万至100万SNP的研究维度，极大地推动了复杂疾病风险定位的研究，遗传分析已经进入了基因组范围关联研究（genome-wide association study，GWAS）阶段。目前，基因组范围关联研究已经应用于40多种复杂疾病，绝大多数研究涉及SNP数目已经超过50万，并通过GWAS成功获得了150多个致病基因。这些疾病基因的获得对于复杂疾病，特别是癌症、糖尿病、心脏病等常见病的研究提供了大量的有用信息，也为进一步揭示这些疾病的发生机制作出了贡献。真正意义上的GWAS开始于2005年前后，应该说，现在还只是它的起步阶段，大规模的GWAs研究还在酝酿，相应的研究策略也在不断的开发。

但正如上文中我们提到的，高维度的SNP数据也给统计学方法带来了很大的压力，多重检验问题困扰着大规模的遗传定位研究。目前，基因组范围关联研究主要通过两个策略来实现风险SNP和风险基因的发现。一方面，采用合并不同实验室样本数据的方法，通过提高研究某个疾病的样本量或SNP密度来加大风险SNP的识别水平，即我们常说的meta分析方法，并且成功应用于乳腺癌、结肠癌和2型糖尿病等研究中。另一方面，采用候选区域精细定位的方法，在较低样本量情况下采用基因组范围关联分析获得候选风险区域，缩小范围后对候选区域加大样本量，进行精细的SNP分型，采用多轮重复策略，最终获得高显著、高精确度的风险位点（图9-10）。这些策略的实施为发现真实的风险SNP提供了可靠的保障，但依然存在花费大、效率低的缺点。

在这样的情况下，人们逐渐将目光从统计方法研究和提高统计显著性角度转移到关联分析结果的信息挖掘上，称之为第二代关联分析策略。第二代关联分析策略将关联分析作为疾病风险权重，期望借助于已知的通路、网络、互作、功能等知识进行位点和基因层面之外的更高层次的信息发现。

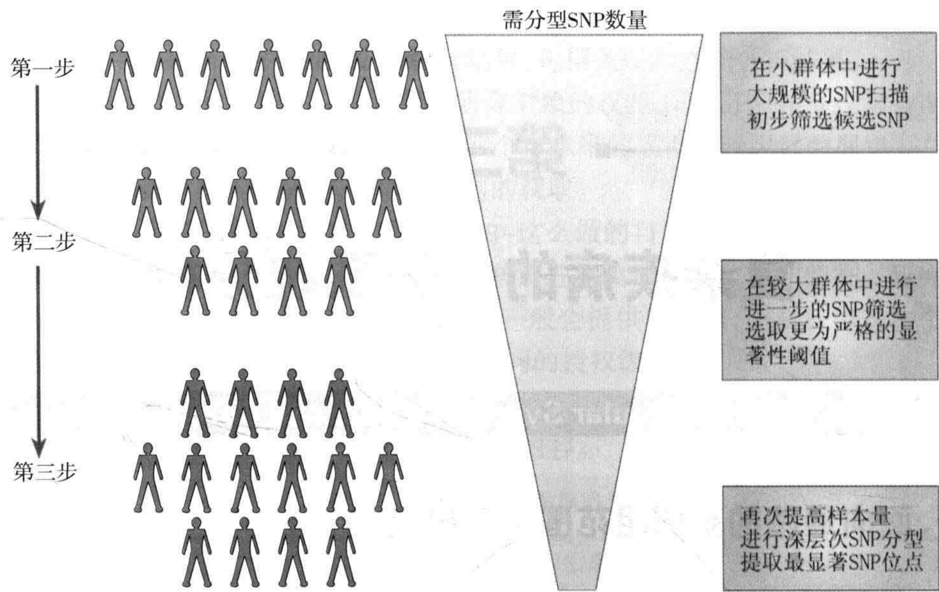


图9-10 精确定位策略提高关联分析可靠性

GWAS研究从全基因组角度进行风险SNP的筛选，实际上也是一定单位点关联分析方法，在一定程度上不足以反映真实的复杂疾病信息，获得的结果一方面存在很高的假阳性率，另一方面不同实验之间还存在低重复率的问题。这一问题在meta分析中有一定的改观，当然也有很多学者试图在基因组范围SNP基因型数据中运用多位点的分析方法，但往往由于极高的计算复杂度而很难获得预期的效果。图9-11显示了一种基于基因组关联分析方法和生物学网络上下文的风险通路优化算法。①计算单位点SNP与疾病风险值，并将基因上最显著的SNP关联p值的负对数作为该基因对疾病的风险值；②将目前已知的人类生理通路中不同路径和位置基因进行网络加权；③以加权的生物学通路作为背景信息，将带关联权重的基因映射到通路中，并计算全通路与疾病之间的风险值，进行风险通路优化和排秩。这种方法实际上考虑到了复杂疾病本身存在的复杂疾病的多基因性、致病基因微效性和致病基因之间的相互作用，利用一种替代的方法进行复杂疾病的多位点分析，在一定程度上避免了多重检验校正和影响，提高了实验的重复率，同时也有利于从通路的角度深入的了解复杂疾病病因学，高效发现疾病关联基因。当然，这种方法也受到现有已知通路信息不全，依赖于单位点关联分析方法存在的随机性影响。

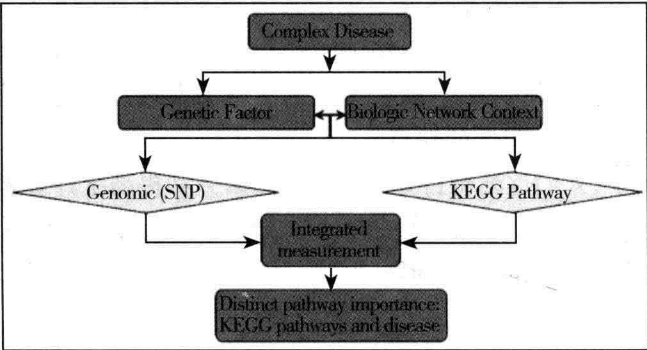


图9-11 基于关联分析的复杂疾病风险通路优化方法

作为一种更为优化的方法，可以从已知的蛋白质互作网络信息出发，来构建一个更为完善的基因与基因之间的先验互作集（潜在的通路背景基因集），将单基因上多个SNP与疾病之间的关联性赋予基因作为基因与疾病之间的风险性，并将基因风险映射到互作网络中，利用相应的网络分析方法，对原有的开放性网络加入组织特异性、细胞定位、生物功能等限制性条件，进行更为可靠的疾病相关的子网筛选。同时，也可以充分利用现有的单通量生物实验获得的成果，加入先验疾病研究基因集，诱导疾病相关子网的提取。这可以有效的利用人们对疾病、人类基因互作关系及其他成熟的研究结果，进行比较全面的复杂疾病潜在疾病通路的发现。如果能够进一步引入代谢子、遗传调控等因素，有可能会从更为科学的角度提升基于SNP研究复杂疾病的可靠性。

这样的策略不仅坚持了疾病基因层面的发现，同时获得的结果还能够从细胞过程和机制的角度来解释疾病的发生，相比原有传统关联分析方法，有着不言而喻的优势。但由于作为研究基础的高通量先验知识本身还存在不完整和假阳性，因此第二代关联分析策略还处于起步和摸索阶段，更为系统的方法研究还存在很大的空间和应用价值。

二、表型性状的分子遗传学

人与人之间形态、生理指标、行为及疾病易感等表型差异共同构成人类本身的多样性。而这些复杂表型的变化往往是由潜在的多位点遗传复杂性，及其遗传等位与个体所处环境之间的不同反应造成的。从DNA变异与表型差异之间的相关性研究的角度，讨论数量或复杂性状产生的原因对于预测疾病发病风险和个性化治疗有重要的意义。这里，我们将与某些数量或复杂性状形成相关的DNA区域称为决定这个性状的数量性状位点（quantitative trait loci，QTL）。早在20世纪早期，人们就开始了对数量性状的研究，并采用遗传多态标记与QTL连锁分析的思想对数量性状进行遗传定位。到20世纪80年代，数量性状研究得到了空前的发展，但是遗传多态标记的缺乏大大限制了它的进一步发展。直到最近几年，随着测序技术的发明和人类单体型计划的实施和完成，大量遗传标记被发现，而且分型成本不断降低，基因组范围数量性状的QTL定位研究迅速发展，并广泛应用于人类性状和疾病研究领域。

经过二十多年的努力，我们已经能够从候选基因、不同遗传背景下的等位分离、生态与环境对表型影响、功能等位效应的分子基础、群体致病等位频率等方面对遗传变异与数量性状形成之进行解释。某些研究通过QTL定位发现了新的疾病或复杂性状位点，并为揭示疾病生物学机制提供新的视野，但明确指出导致表型和疾病形成的变异，只占全部表型决定子的一小部分，通过QTL定位直接发现表型相关的基因更是少之又少。不过，这一情况并不取决于目前对QTL定位的研究方法，而是与现在的DNA和RNA的测序水平相关的，将会伴随新的高通量、快速、低廉的测序技术的产生而取得新的突破。

与质量性状相比，数量性状的遗传研究要困难得多，主要是由于质量性状可以通过表型来辨别，而数量性状表型上的差异不明显，基因型与表型间难以找到准确的对应关系。而由于人类群体不可能像在动植物中进行杂交实验，所以对人类群体的数量性状定位更加困难。无论是在人类还是在动植物中，数量性状定位的基本原理都是数量性状位点与可见的分型分子标记之间存在遗传连锁。如果某个QTL与某个分子标记（SNP）相联系，在此位点上具有不同等位的个体具有不同的数量性状平均值。基于这样的思考，在人类中我们虽然不能

进行特定的位点杂交实验，但是可以通过遗体遗传学方法进行位点与数量表型均值之间的相关性检验，从而完成数量性状定位。常用的数量性状定位分子标记除SNP外，还有插入/删除多态、微卫星等。

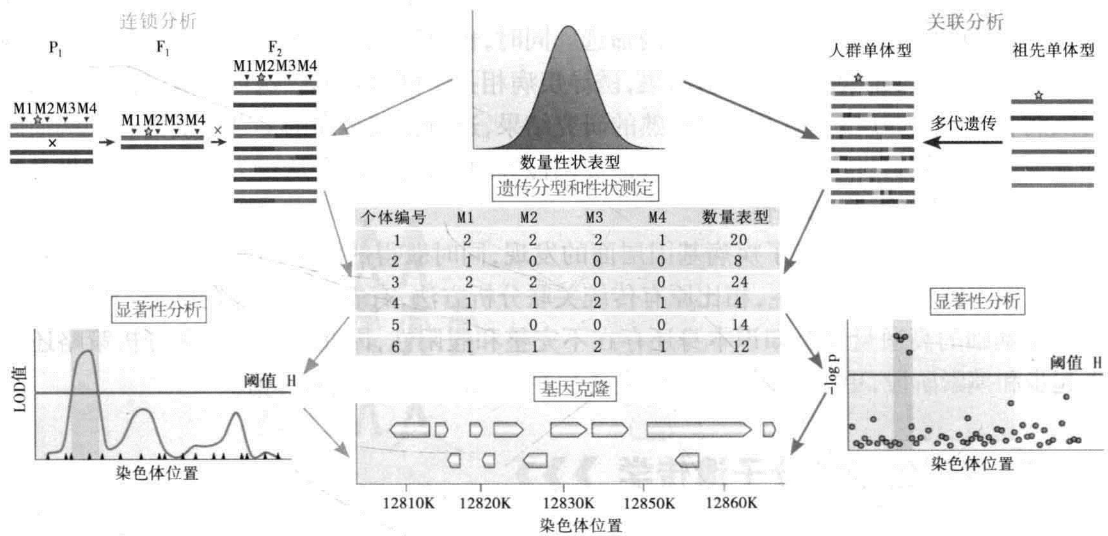


图9-12　数量性状分析流程图

显著相关位点的检测和原因基因克隆是数量性状定位的两个要点。图9-12显示了基于SNP的QTL分析基本过程。在人类样本分析中，由于家系信息难以获得，而主要通过关联分析的方法进行检验（图9-12右侧），而相对的，动植物研究中可以方便地进行杂交实验，一般采用连锁分析的方法（图9-12左侧）。

人类遗传学中进行数量性状定位最常用的方法是线性回归和方差分析。方差分析进行数量性状定位类似于自由度为2的皮尔森检验，这里，将0假设定义为数量性状与SNP基因型没有相关性，备选假设为有相关性。而线性回归方法用于数量性状研究主要考虑SNP基因型与数量性状平均值之间的关系，自由度为1。两种情况下均要求数量性状呈近似正态分布，如果分布有偏差，可以考虑进行对数转换。

三、复杂疾病的系统遗传学 》》》

对于常见的数量性状，我们可以很自然地联系到身高、体重等看得见的研究对象，除此之外，还能够想到血压水平、血糖水平等与人体健康检查有关的症状反映或生化指标。随着SNP分型技术和基因表达作谱技术的发展，越来越多的研究把目标锁定在人类基因表达调控子的发现上，通过对同质性样本的SNP分型及基因表达绘制图谱，期望建立分子标记与基因表达之间的联系，这一过程被称作表达数量性状位点（expression quantitative trait loci，eQTL）定位。2002年，Rachel等人利用芯片表达技术与关联分析相结合的方法研究杂交酵母的遗传变异与基因表达之间的相关性。发现众多的基因表达是受遗传变异影响的，不同的遗传等位有可能导致不同的表达效果。这为从遗传角度揭示表型形成提供了一个有利的证据。2004年，Michael等人将eQTL研究引入到人类基因组研究领域，通过对14个家系196个个体的2980个SNP与3553个转录子

表达的测定，发现170多个SNP与之邻近的基因表达之间存在相关性，并把这种关系称之为Cis关联，此外，还存在众多的远距离的SNP与基因表达之间的相关性（Trans），这一研究证实人类基因表达受到了广泛的遗传调控，而且可以通过数量性状定位的方法将遗传变异和表达进行关联。

目前，与基因组范围关联研究发展相适应，eQTL研究已经从最初的数以千计的SNP与基因表达规模发展到数以十万计的SNP和2万多基因表达之间的关系，而且从基于家系和模式生物的研究逐渐过渡到基于不相关个体的研究，发现的人类遗传与表达之间的关系也越来越多。2007年10月，《Nature Genetics》连续发表3篇文章进行人类基因组范围的eQTL研究。Barbara等人基于HapMap样本，进一步测定14 000个基因的表达情况，进行了四个群体的eQTL研究，从群体比较方面揭示遗传变异调控基因表达的群体差异性。Harald等人将基于淋巴细胞的研究样本量提高到1240个个体，研究的基因数高达1.9万。而Anna等人的研究首次将疾病因素引入到基因组范围关联研究中，通过研究哮喘家系中的遗传变异与基因表达之间的关系，提出可能实现联合eQTL与疾病的研究，易化关联研究中的功能元件提取。2008年3月Valur等人联合基因表达、遗传变异及临床肥胖指标进行合并的QTL研究进行疾病相关的遗传子及功能元件识别，并在此基础上提出从分子网络的角度研究复杂疾病。

表达数量性状定位的提出为生物医学研究展开了更为广阔的视野，也为从DNA→表达→分子表型→性状的研究提供了可能。在这样的背景下，科学家们提出系统遗传学（systems genetics）概念，即希望从全面的生物学资源出发，研究遗传因素对人体生理病理的影响。Trudy等人在此基础上提出未来的遗传定位研究的着眼点（图9-13），期望借助系统遗传学工具实现从分子到整体的全面了解。而表达与标记、表达与表达之间的相关又能向网络的层面进行转化，对于获取生理学或病理学功能信息将产生直接而有效的影响。

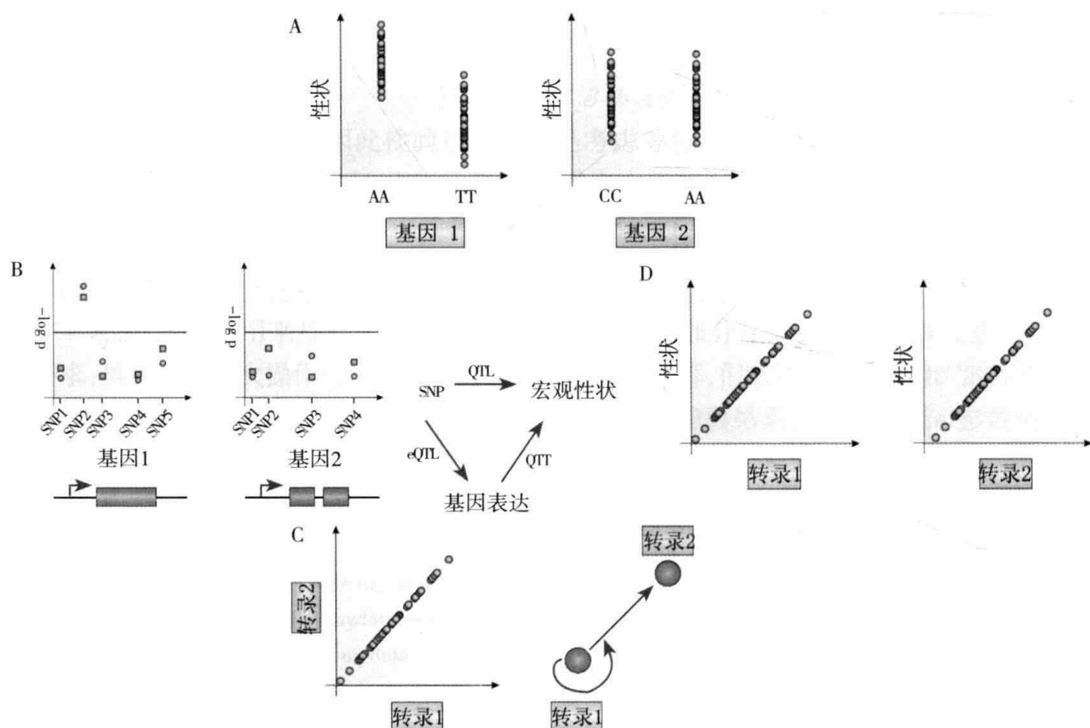


图9-13　系统遗传学思想构建遗传调控网络揭示性状形成机制

图9-13A中显示了广泛应用的纯合基因型与数量性状平均值线性回归方法获得与性状相关的遗传标记（SNP）的过程，即我们常说的QTL定位，将一个宏观的指标与分子层面的标记相互联系。B图展示的是基因表达与SNP之间的关联分析过程，即上文讲到的eQTL定位，从而捕获影响基因表达的SNP，这些SNP我们称为调控SNP，这个过程将基因表达和SNP进行关联。C图进行的是基因共表达分析。通过B、C两图可以构建出一个基于遗传分析的调控网络，而由于我们已经在SNP与表型之间建立起联系，借助分子网络的分析手段，能够指导我们发现影响性状形成的基因集，甚至指导我们发现与性状发生密切相关的网络模块或通路。目前，由于同时进行基因型、基因表达、人类表型的测定和收集过程耗时、耗力，而且花费巨大，在一定程度上限制了系统遗传学的开展，但随着技术的革新，这样的一种研究思想将逐步成为人类生理、病理研究的必由之路。

第四节

系统遗传学集成软件工具

Section 4 Important Tools in SNP Studies

一、Plink软件包与基因互作

PLINK(http: //pngu.mgh.harvard.edu/~purcell/plink/contact.shtml#cite)是一个免费、开源的全基因组关联分析工具集，旨在用有效地计算方式进 行常规的及大规模的遗传分析。PLINK的主要功能包括: 数据处理和统计描述、群体分层检测、关联分析、IBD估计及上位效用检测。PLINK一般只适用于群体数据，不适用于家系数据。本节中以上位效应为例介绍PLINK。

PLINK用于检测SNP-SNP间上位效应所用的默认检验模型主要有线性回归和罗杰斯特回归两种，取决于表型是数量性状还是二值性状。基于每一个A和B的等位基因情况，建立一个模型:

$$Y \sim b_0+b_1A+b_2B+b_3AB+e \quad (9-4)$$

互作检验基于系数b_3，因此检验过程中只是考虑等位基因之间的上位效用，不考虑协变量。SNP-SNP上位效应检验可以在病例/对照样本中进行，也可以只在疾病样本检测(也叫做case-only)。在病例/对照样本中检测SNPxSNP上位效应，用以下命令:

```
plink --file mydaya --epistasis
```

--epistasis命令用来检验大量的SNP-SNP互作，但大部分互作没有显著意义或不符合用户要求，虽然可能一次操作会进行数百万或数十亿行的计算，但默认只输出$p < 10^{-4}$的互作，或者用--epi1参数设定。如果数据集比较小，期望输出所有的检验结果，可以用--epi1参数测定，如:

```
plink --file mydata --epistasis --epi1 0.0001
```

同时也可以通过命令设定进行检验的SNP集合，相应的模式如下:

```
任意SNP之间：plink --file mydata --epistasis
集合1内部：plink --file mydata --epistasis --set-test --set epi. set
集合1-全部：plink --file mydata --epistasis --set-test --set epi. set --set-by-all
集合1-集合2：plink --file mydata --epistasis --set-test --set epi. set
```

epi.set可以只含有一个数据集，也可以包含有多个数据集。对于每一个数据集开头有数据集名称，数据结尾有END符号。

在病例样本中检测SNP-SNP上位效应，有两种近似但更快速的参数命令：--fast-epistasis 和--case-only，用以下命令执行：

```
plink --file mydata --fast-epistasis --case-only
```

目前，在case-only分析中，默认状态下只考虑距离1Mb以上或不在同一条染色体上的SNP，其他SNP上位效应计算，可以通过-gap参数设定SNP之间的距离，如下：

```
plink --file mydata --fast-epistasis --case-only --gap 5000
```

-gap是一个很重要的参数，但使用时应当慎重，因为用case-only检验上位效应的两个SNPs在群体中应处于连锁平衡状态。

通过以上的命令，我们已经了解plink计算上位效应的基本方法，表9-4中列出了计算上位效应中常用的命令和默认参数。

表9-4 Plink计算上位效应的参数列表

命令	参数或缺省值	描述
--file		指定.ped和.map文件
--epistasis		进行SNP之间的上位效应分析
--fast-epistasis		快速进行任意两个SNP之间的上位效应计算
--twolocus	SNP SNP	显示两个SNP互作列表
--case-only		只能疾病样本进行上位效应计算
--gap	1000	限定距离的SNP上位效应计算
--epi1	0.0001	输出上位效应计算*p*值小于阈值的对
--set-by-all		检验集合中的SNP上位效应
--nop		进行快速筛选，不计算*p*值
--genepi		进行基于基因的上位效应计算

Plink的输入文件有两个，分别以.ped和.map作为后缀。PED和MAP文件是用空格或Tab分割的文件，PED文件的每一行代表一个样本描述，并且前六列描述信息是必需的，如缺失应当用0代替，但必须含有表型信息。MAP文件的每一行是一个SNP的染色体定位（表9-5）。

表9-5 PED和MAP文件说明

列数	PED文件	MAP文件
第1列	个体所在家系ID[a]	SNP所在染色体[b]
第2列	个体在家系中的编号	SNP标识符
第3列	个体对应的父亲编号	SNP的遗传距离
第4列	个体对应的母亲编号	SNP的绝对位置
第5列	性别	
第6列	表型状态[c]	

注：a1代表男性，2代表女性，其他标记表明性别未知，b分别使用1-22数字，X，Y表，0代表所在染色体未知，c1表示为对照个体，2表示为疾病个体。

输出文件包括plink.epi.cc和plink.epi.cc.summary。plink.epi.cc文件显示以下信息：①CHR1，第一个SNP所在的染色体；②SNP1，第一个SNP识别符；③CHR2，互作的SNP所在的染色体；④SNP2，互作的SNP识别符；⑤OR_INT，两位点互作的ood ratio值；⑥STAT，自由度为1的卡方检验统计量；⑦P，显著性水平。plink.epi.cc.summary文件显示以下信息：①CHR，染色体编号；②SNP，SNP标识符；③N_SIG，上位效应的显著性检验（p<="--epi2"阈值）；④N_TOT，可执行检验；⑤PROP，可执行检验的百分数；⑥BEST_CHISQ，与互作的SNP检验结果；⑦BEST_CHR，互作的SNP染色体编号；⑧BEST_SNP，互作的SNP标识符。

二、基因组范围关联研究软件包SNPtest

SNPtest（http://www.stats.ox.ac.uk/%7Emarchini/software/gwas/snptest.html）是一个强大的基因组范围关联研究软件包，它可以对单个SNP关联进行频率检验或贝叶斯检验，值得注意的是，它的实施只适合于二进制（病例对照）性状，但该软件可以根据任意的协变量集进行设置，并且能够考虑基因型的不确定情况。目前，被广泛应用的WTCCC中，7套复杂疾病的基因组范围关联研究，就是采用该软件进行的数据分析。SNPtest同时提供了2000个个体中100个SNP的疾病-对照示例文件。

（一）软件的输入文件

SNPtest允许分析多组个体。每组数据存为两个文件：第一个文件为样本文件，存储的是ID号、关联协变量和每组个体的表型信息；第二个文件为基因型文件，存储的是每组基因型数据。软件当中包括的例子数据集中每组的样本和基因型文件分别有符合要求的_sample和_gen样文件。

基因型文件格式（_gen）：该文件每行表示一个SNP信息，前5列分别为：SNP ID、RS ID、SNP碱基对位置、两个等位基因（M、N）；接下来的3个数字表示三种基因型MM、MN、NN在第一个个体中出现的概率值，再接下来的3个数字表示三种基因型在第二个个体中出现的概率，以此类推。并且个体的顺序应该与_sample文件中个体的顺序相同。同时，考虑到缺失基因型情况，因此基因型概率之和不必均为1。当对多组执行SNPtest时，我们假设每组数据的SNP集大小相同并且这些SNP在每组的基因型文件中的存储顺序相同。

样本文件格式（_sample）：该文件包括三个部分：第一行，表示每一列的名字；第二行，每一列所存储变量的类型；接下来的每行表示一个个体的详细相关信息。例如：

ID_1	ID_2	missing	cov_1	cov_2	cov_3	cov_4	phenotype_1
0	0	0	1	2	3	3	p
1	1	0.007	1	2	0.0019	−0.008	1.233
2	2	0.009	1	2	0.0022	−0.001	6.234
3	3	0.005	1	2	0.0025	0.0028	6.121
4	4	0.007	2	1	0.0017	−0.011	3.234
5	5	0.004	3	2	−0.012	0.0236	2.786

第一行分别表示：个体的第一个ID号、第二个ID号、个体中缺失值的比例，这三个是必须

要有的，接下来的分别表示变量的名字。上面的例子中，有4个协变量cov_1，cov_2，cov_3，cov_4和1个表型名字phenotype_1.。

第二行表示每列中变量的类型，前3个设置为0，接下来的位置应遵循下面的规则：

1	离散的协变量（用正整数表示），对关联进行Mantel-Haentzel检验
2	离散的协变量（用正整数表示），对关联进行跨群体的整合检验
3	连续协变量
p	表型

（二）软件包中的分析模块

1. 数据的统计描述　SNPTEST最基本的用途是对SNP数据基本信息进行描述，生成包括基因型数目、等位基因频率、SNP缺失数据比例和优势比等的描述信息，这个功能用以下命令行可以实现：

```
. /snptest –cases . /example/cases. gen . /example/cases. sample
–controls . /example/controls. gen . /example/controls. sample –o . /example/ex. out
```

2. 压缩文件输入命令　SNPtest支持压缩文件，当基因型文件相当大的时候会以压缩文件的形式给出，那么在SNPTEST中有一个命令-gen_gz就表示输入的文件为压缩形式，在命令行中输入：

```
. /snptest –gen_gz –cases. /example/cases. gen. gz. /example/cases. sample
–controls. /example/controls. gen. gz. /example/controls. sample –o . /example/ex. out
```

会输出同上面所介绍的相同的结果文件./example/ex.out。

3. 计算数据缺失率　样本文件的第三列包含每个个体的缺失数据比例。这有利于滤除那些缺失数据率高的个体。命令-create_misske可用来计算形成样本文件所需的缺失数据率。例如，计算第一个对照组的缺失数据率，可以使用如下的命令行：

```
. /snptest –create_miss. /example/controls. gen –o. /example/ex. out
```

4. 排除SNP及（或）个体　排除SNP：命令-exclude_snps可被用来指定一个文件，该文件中包含一列分析当中应当排除的SNP。例如，文件./example/snps.list包含了一列example文件数据的前10个SNP编号，为了从分析当中排除这些SNP我们使用下面的命令行：

```
. /snptest –cases. /example/cases.gen. /example/cases. sample
–controls. /example/controls. gen. /example/controls. sample –o. /example/ex. out
–exclude_snps. /example/snps. list
```

另外，程序还提供命令-snpid来对单个的SNP执行此功能。例如，对编号为61的SNP运行SNPTEST，我们用下面的命令行：

```
. /snptest –cases. /example/cases. gen. /example/cases. sample
–controls. /example/controls. gen ./example/controls. sample –o. /example/ex. out –snpid 61
```

排除个体：命令-exclude_samples可被用来指定一个文件，该文件中包含一列分析当中应当排除的个体。例如，文件./example/samples.list包含了一列样本数据文件的前10个个体的ID号，为了从分析当中排除这些个体我们用下面的命令行：

```
. /snptest -cases. /example/cases. gen. /example/cases. sample
-controls. /example/controls. gen. /example/controls. sample -o. /example/ex. out -exclude_
samples. /example/samples. list
```

命令-miss_thresh可被用于排除那些缺失数据比例达到某一水平的个体。例如，为了指定最大缺失数据比例为1%，我们利用下面的命令行：

```
. /snptest -cases. /example/cases. gen. /example/cases. sample
-controls. /example/controls. gen. /example/controls. sample -o. /example/ex. out -miss_
thresh 0.01
```

5. 哈代温伯格平衡检验　命令-hwe表示在输出结果中显示出每个SNP的HWE检验结果。例如：

```
. /snptest -cases. /example/cases. gen. /example/cases. sample
-controls. /example/controls. gen. /example/controls. sample -o. /example/ex. out -hwe
```

将产生一个输出文件./example/ex.out，该文件的列包含的是对每个对照组的精确HWE检验的p值、对照组的整合集、每个病例组HWE检验p值、病例组整合集。

6. 基本的关联检验　病例对照检验：对加性、显性、隐性、常规及杂合子5个模型的关联进行标准频率病例对照检验，可由命令-frequentist来执行。例如，下面的命令行被用来对这四种模型进行检验：

```
. /snptest -cases. /example/cases. gen. /example/cases. sample
-controls. /example/controls.gen. /example/controls. sample -o. /example/ex. out -frequentist
1 2 3 4 5
```

5种不同的模型编号为：1-加性模型，2-显性模型，3-隐性模型，4-常规模型，5-杂合子模型。加性模型是对加性遗传效应进行Cochran-Armitage检测。显性模型和隐性模型是将AA基因型当作起点基因型。常规模型则是对关联进行自由度为2的标准检验。

输出文件为./example/ex.out，包含了每个SNP所有如前面描述的概要信息。四个检验的p值分别在frequentist_add，frequentist_dom，frequentist_rec，frequentist_gen and frequentist_het列中给出。

数量性状检验：对SNP与一个数量性状关联的检验可以用-qt命令来执行。对每个SNP的关联该命令是通过F-检验来执行的。命令-frequentist被用来指定每个SNP的基因型编码。每个个体的基因型必须出现在样本文件当中。在默认情况下，检验将使用样本文件当中的第一个基因型。用户应当用-pheno这一命令来指定你所要检测的表型。例如下面的命令行，是对例子数据集中的第二个表型在5个不同模型中进行检验：

```
. /snptest -cases. /example/cases. gen. /example/cases.sample
-controls. /example/controls.gen. /example/controls. sample -o. /example/ex. out -qt -pheno
2 -frequentist 1 2 3 4 5
```

7. 协变量 命令–cov被用于存在协变量时进行关联的检测。例如,在考虑到样本文件中的第二个协变量时对一个加性模型进行关联检测时,我们可用下面的命令行:

```
. /snptest –cases. /example/cases. gen. /example/cases. sample
–controls. /example/controls. gen. /example/controls. sample –o. /example/ex. out –frequentist
1 –cov 2
```

产生输出文件./example/ex.out,该文件包含了一个表头frequentist_add_cov_1的列,包含了基于协变量检测得到的p值。

8. 贝叶斯检验 用命令–bayesian可对5个标准遗传模型进行贝叶斯检验。例如,下面的命令行:

```
. /snptest –cases. /example/cases. gen. /example/cases. sample
–controls. /example/controls. gen. /example/controls. sample –o. /example/ex. out –bayesian 1
2 3 4 5
```

产生一个输出文件,包含以下几列信息: bayesian_add, bayesian_dom, bayesian_rec, bayesian_gen and bayesian_het。

9. 考虑基因型不确定的情况 改变所需阈值: 程序的默认设置为利用一个阈值来将基因型命名为AA、AB、BB或NULL。对关联做频率和贝叶斯检验将基于这些默认阈值基因型来执行。如果基因型大于所需阈值那么我们选择最大概率的基因型,否则引入NULL基因型。默认阈值为0.9,此阈值是可改变的,用命令–call_thresh来改变。例如,为了产生一个基于阈值为0.95的基本检验集,我们执行下面的命令行:

```
. /snptest–cases. /example/cases. gen. /example/cases. sample
–controls. /example/controls. gen. /example/controls. sample –o. /example/ex. out –call_
thresh 0.95 –frequentist 1 2 3 4
```

频率检验: 命令–proper可被用来考虑基因型不确定的情况。该命令执行一个基于缺失数据可能性的统计学检验。该命令与–cov、–qt两个命令一起用。例如,下面的命令行:

```
. /snptest –cases. /example/cases. gen. /example/cases. sample
–controls. /example/controls. gen. /example/controls. sample –o. /example/ex. out –frequentist
1 –proper
```

此命令产生一个输出文件,包含以下几列信息: frequentist_add_proper、frequentist_add_proper_info、frequentist_add_proper_beta_1、frequentist_add_proper_se_1。

贝叶斯检验: 在考虑基因型不确定的情况下计算贝叶斯因子时,是通过基于基因型概率抽取基因型使贝斯因子结果平均化。命令–nsamp指定了应该用到的基因型样本数目。–certainty_thresh命令可用来指定在哪个SNP中进行了抽样。例如下面的命令行:

```
. /snptest –cases. /example/cases. gen. /example/cases. sample
–controls. /example/controls. gen. /example/controls. sample –o. /example/ex. out –bayesian 1
–nsamp 100 –certainty_thresh 0.95
```

此命令产生了一个列为bayesian_add_samp的输出文件,表示在加性模型中样本均值为log 10贝叶斯因子。

三、连锁分析和数量性状分析工具Merlin

Merlin(http: //www.sph.umich.edu/csg/abecasis/Merlin/index.html)是一个利用稀疏遗传树进行系谱分析的软件包。Merlin利用稀疏树来代表系谱中的基因,它是最快的谱系分析软件包之一。Merlin能够被用于参数或非参数的连锁分析,以回归为基础的连锁分析或对数量性状的关联分析, IBD和亲属关系的估计,单体型分析,错误检测和模拟分析。在大部分分析中标记之间可以存在连锁不平衡状态,并且能够比其他的系谱分析软件包处理更多的标记。

Merlin进行普遍的家系分析。输入文件描述数据集中个体之间的关系,储存了标记基因型,疾病的状况和数量性状标记信息,并提供了位点定位及等位基因频率信息。Merlin支持QTDT或LINKAGE格式的输入文件。这两种格式非常相似,在以下的讨论中我们将主要关注QTDT格式。

(一) 群体分层分析

虽然家系会变得非常复杂,在一个家系文件中所有用于重建个体间关系的信息可以概括为5个项目: 一个家庭的标识符,个体识别码,与每位家长的链接(如果有的话),最后一个指标是每个个体的性别。

以下为是一个虚拟的家系文件:

FAMILY	PERSON	FATHER	MOTHER	SEX
example	granpa	unknown	unknown	m
example	granny	unknown	unknown	f
example	father	unknown	unknown	m
example	mother	granpa	granny	f
example	sister	father	mother	f
example	brother	father	mother	m

这些关键值构成了任何一个家系文件的前五列。由于在早期的遗传程序中存在的限制,文本标识符通常被唯一的数值所取代。每个标识符被唯一的整数所替代且将性别编码为女性为2,男性为1之后,一个基本的以空格分隔的家系文件会是以下这种形式:

```
<contents of basic. ped>
1   1   0   0   1
1   2   0   0   2
1   3   0   0   1
1   4   1   2   2
1   5   3   4   2
1   6   3   3   1
<end of basic. ped>
```

一个家系文件可以包括多个家庭。每个家庭都有唯一的结构,在数据集中与其他家庭之间存在独立性。

(二) 表型与基因型

通常标准的5列之后的各种类型的基因数据,包括离散的表型数据,数量性状数据和标记基因型数据。

疾病状况通常在单独的一列进行编码:

```
U or 1 for unaffecteds, A or 2 for affecteds, and X or 0 for missing phenotypes.
```

编码数量性状时用X表示缺失值(也可以使用一种特殊的数值表示缺失的表型值,但该程序容易出错,不推荐)。

标记基因型被编码成用两个连续的整数,对于每一个等位基因用一个"/"进行分隔,或自1.1版本后使用字母"A","C","T"和"G"来编码。为了表示缺失的基因,可以用0,X或N。以下是所有有效的基因型项1/1(等位基因为1的纯合子),0/0(缺失的基因型),及3/4(等位基因为3和4的杂合子)。在Merlin的较新版本A/A,A/C和C/C也是有效的基因型。对于X染色体,男性应该像他们好像有两个相同的等位基因那样被编码。

以下为前面的家系文件添加了疾病状况,对数量性状的测量值和两个标记的基因型后所呈现的形式:

```
<contents of basic2. ped>
        1   1   0   0   1   1   x       3   3   x   x
        1   2   0   0   2   1   x       4   4   x   x
        1   3   0   0   1   1   x       1   2   x   x
        1   4   1   2   2   1   x       4   3   x   x
        1   5   3   4   2   2   1.234   1   3   2   2
        1   6   3   4   1   2   4.321   2   4   2   2
<end of basic2. ped>
```

注意第5个个体和第6个个体,她们都被标记成易感(她们在第6列的值为2),其他的每个个体都被标记成非易感的(他们在第6列的值为1)。她们的数量性状(第7列)值为1.234和4.321。尽管每个个体在第一个标记上都进行了基因分型,但对于第二个标记,只有个体5和个体6进行了基因分型。

(三) 家系数据分析

家系文件所包含的标记基因型,疾病的状况和数量性状变量的个数只受可用内存的限制。由于每个家系文件具有唯一的结构(除了第一个5列),其内容必须在与其配对的数据文件中被描述。

数据文件包括家系文件中的每行数据项,显示出了数据类型(将标记编码为M,将易感状况编码为A,将数量性状编码为T,并将相关变量编码为C)并为每一个数据项提供了一个用一个单词表示的标签。对应于上述家系的包含有一个易感状况,接下来是一个数量性状和两种标记基因型的数据文件的具体形式如下所见:

```
<contents of basic2. dat>
        A        some_disease
        T        some_trait
        M        some_marker
        M        another_marker
<end of basic2. dat>
```

可以利用pedstats(包含在Merlin分布中)得到任何一组家系文件和数据文件的概括性描述。要运行pedstats你必须提供你的数据文件的名称(-d命令行选项)和家系文件的名称(-p命令行选项)。在Merlin的例子的目录中,尝试下面的命令:

```
prompt> pedstats -d basic2. dat -p basic2. ped
```

小提示:在Merlin和Pedstats的新版本中,就可以组合多个家系和数据文件。这种方法在分析多个不同的子集或你想通过染色体或区域划分基因型时非常方便。例如,如果你的表型数据存储在pheno.dat和pheno.ped文件中,且你的基因型数据存储在geno.dat和geno.ped文件中,你可以利用以下命令行组合它们:

```
prompt> pedstats -d pheno. dat, geno. dat -p pheno. ped, geno. ped
```

(四) 遗传定位

为了分析遗传标记,Merlin需要它们在染色体上的定位信息。这通常提供了一个定位文件。如果你正在使用性别平均定位,此文件中的每个标记占一行三列,显示出染色体,标记名称和位置(以厘摩为单位)。如果你正在使用的是性别特异性定位,你需要另外两列分别来指定沿女性遗传方向定位的标记位置和沿男性遗传方向定位的标记位置。

数据文件和定位文件可以包含不同的标记集合,但那些在定位文件中缺少标记就会被Merlin忽略。下面是一个典型的定位文件,如下所示:

```
<contents of basic2. map>
        CHROMOSOME        MARKER            POSITION
            24            some_marker        123.4
            24            another_marker     136.2
<end of basic2. map>
```

这里是一个精密版本的定位文件,包括每个标记的性别特异性定位位置:

```
<contents of file with sex-specific map>
  CHROMOSOME    MARKER          POSITION   FEMALE_POSITION   MALE_POSITION
      24        some_marker      123.4         146.8            100.0
      24        another_marker   136.2         166.4            103.0
<end of sex-specific map>
```

使用划分后的数据和定位文件作出了一个非常简单的文件结构,并允许Merlin在一个单一的运行中分析多个染色体。

(五) 等位频率分析

LINKAGE格式数据文件指定在每个位点上的等位基因的个数和它们的频率。当使用QTDT格式输入文件时，Merlin通过计算所有个体的等位基因个数来估计等位基因的频率。如果这种方法得到的等位基因频率对于现在的分析不适合，你需要对等位基因频率进行最大似然估计(–fm 命令行选项)，规定合适的等位基因频率(–fe)，要求只通过在创建者间进行计算所获得的估计值(–ff)，或提供一个自定义等位基因频率文件(– f文件名选项)。

一个自定义等位基因频率文件指出了在每一个标记处的所有标记等位基因的频率。对于每一个标记，用来命名标记的单一的标题行之后接下来是一系列等位基因频率，它可占用很多行。

每个标题行以M作为标签，并包括标记的名称。接下来的一系列等位基因频率有两种可选择的格式: ①经典格式: 等位基因频率列表中的每行以F作为标签，且列表中所有等位基因的频率都是连续的，以等位基因1作为开始。这种格式对于具有少量等位基因的标记来说很方便; ②扩展格式: 等位基因频率列表中的每行以A作为标签，且包含一个数字的等位基因标签，接下来是一个等位基因频率。在列表中没有被明确列出的等位基因被估计成频率为0。

经典等位不平衡模式

例如，如果some_marker有四个等位基因，频率分别为0.1，0.2，0.3和0.4，another_marker有两个等位基因，频率分别为0.6和0.4，那它们在文件中为以下形式:

<contents of basic2. freq>
M some_marker
F 0.1 0.2 0.3 0.4
M another_marker
F 0.6 0.4
<end of basic2. freq>

或

<contents of basic2. freq>
M some_marker
F 0.1
F 0.2
F 0.3
F 0.4
M another_marker
F 0.6
F 0.4
<end of basic2. freq>

(六) 等位扩展

这种格式被推荐用于微卫星和其他具有大量等位基因的标记。例如，如果你正在分析一个具有152个、154个和156个基础对的等位基因的微卫星标记，且它们的频率分别为0.5、0.4和0.1，那么频率文件可以被写成以下的形式:

<contents of allele frequency file>
M some_microsatellite
A 152 0.5
A 154 0.4
A 156 0.1
<end of allele frequency file>

（七）关联分析模块

Merlin也可以检测一个SNP与一个或多个数量性状之间的关联性。在Merlin中进行的关联性检测包括一个集成的基因型推理功能，它可以在一些基因型缺失的情况下提高工作效能。在这个例子中，我们将看到如何利用Merlin进行关联分析，以及如何利用集成的基因型推理功能估计缺失的基因型。

Merlin进行的关联检测可以用来全基因组关联性扫描，或用于候选区域研究。不过，重要的是要注意与标准的以家庭为基础的关联测试的相比，在Merlin中进行的检测并不控制群体分层。如果群体分层是一个要关注的方面，那么群体的成员应该作为相关变量被包括在其中或用基因控制的方法来矫正结果。

要运行Merlin中的关联分析，我们需要指定数据集合（-d参数），一个家系（-p参数）和定位文件（-m参数）。此外，我们需要下列关联性检测之一：打分检测（-fastAssoc）或似然比检验（-assoc）。打分检测（-fastAssoc）能够快速、理想的筛选大量的标记（例如，在一个全基因组范围关联扫描的第一阶段中），而更精确的似然比检验（-assoc）可以用来评估数量较少的标记（例如，可用于在候选区域进行挑选的后续分析中）。在只包含较小家系的数据集或当被评估的影响较小时，这两项检测会给出类似的结果。

```
prompt> merlin -d assoc. dat -p assoc. ped -m assoc. map -fastAssoc
prompt> merlin -d assoc. dat -p assoc. ped -m assoc. map -assoc
```

-assoc和-fastAssoc，是两个最常用于检测关联性的命令，上面的命令行是采用这两个命令的输入格式。这些命令在Merlin中用于常染色体分析，且在Minx中用于X-连锁标记分析。命令运行中，还可以采用-PDF选项和-inverseNormal选项对结果进行了图形化的概括或自动变换性状使它们遵循平稳的正态分布。

（徐良德 李晋 陈曦）

参考文献

1. *A haplotype map of the human genome.* Nature, 2005. 437(7063): 1299-1320.
2. Balding D. J. *A tutorial on statistical methods for population association studies.* Nat Rev Genet, 2006. 7(10): 781-791.
3. Carlson C. S, Eberle M. A, Kruglyak L, et al. *Mapping complex disease loci in whole-genome association studies.* Nature, 2004. 429(6990): 446-452.
4. Conrad D. F, Jakobsson M, Coop G, et al. *A worldwide survey of haplotype variation and linkage disequilibrium in the human genome.* Nat Genet, 2006. 38(11): 1251-1260.
5. Cookson W, Liang L, Abecasis G, et al. *Mapping complex disease traits with global gene expression.* Nat Rev Genet, 2009. 10(3): 184-194.
6. Hirschhorn J. N, Daly M. J. *Genome-wide association studies for common diseases and complex traits.* Nat Rev Genet, 2005. 6(2): 95-108.
7. Jakobsson M, Scholz S. W, Scheet P, et al. *Genotype, haplotype and copy-number variation in worldwide human populations.* Nature, 2008. 451(7181): 998-1003.

8. Mackay T. F, Stone E. A, Ayroles J. F. *The genetics of quantitative traits: challenges and prospects.* Nat Rev Genet, 2009. 10(8): 565–577.

9. McCarthy M. I, Abecasis G. R, Cardon L. R, et al. *Genome-wide association studies for complex traits: consensus, uncertainty and challenges.* Nat Rev Genet, 2008. 9(5): 356–369.

10. Rockman, M. V, Kruglyak L., *Genetics of global gene expression.* Nat Rev Genet, 2006. 7(11): 862–872.

第十章
CHAPTER 10

miRNA与复杂疾病
miRNA AND COMPLEX DISEASE

第一节

miRNA与其靶基因

Section 1 miRNA and Target Gene

一、miRNA概述

1993年，研究人员Victor Ambros等人发现了一个能够影响秀丽新小杆线虫发育的基因。他们发现其产物是一个小的非编码RNA，将其命名为Lin-4。它通过与基因*Lin-14*的3'端非翻译区（untranslated region，UTR）相互作用，调节线虫的发育。在2000年，另一研究小组在对秀丽新小杆线虫发育过程的研究中又发现了另一个小的调控RNA——Let-7。短期内，研究人员在哺乳动物中发现许多Let-7的同源物。这些同源物与线虫中的Let-7具有相似的时间表达模式。不久后，人们在线虫、果蝇、斑马鱼、拟南芥、水稻以及人类等多种真核生物中找到了上百个类似的小分子RNA，并将其称为miRNA。

miRNA是一种长度大约为22nt的内源性单链RNA分子，能够调控基因的表达。据推测，人类有超过三分之一的基因受miRNA调控。目前，已有超过1000种人类miRNA被发现。随着人们对miRNA研究的深入，许多其他类型的小RNA陆续在动物、植物以及真菌中被发现。这些小RNA包括内源性小干扰RNA（small interfering RNA，siRNA）和piwi-interacting RNA（piRNA）。同miRNA一样，这些小RNA具有RNA沉默的功能。然而，miRNA与这些小RNA在生物合成上明显不同。miRNA是来自于自身转录本所形成的发夹结构，而其他类型的小RNA或者来自于更长的发夹结构，或者来自于RNA二聚物（siRNA），还可能来自没有任何双链结构的前体（piRNA）。

人们对通过miRNA基因组的分析发现，超过50%的哺乳动物miRNA位于基因内，这些miRNA可以与它们的宿主基因一同转录。其他的miRNA则位于基因间区，一般认为，这些miRNA具有自己独立的启动子，可以形成独立的转录单元。编码miRNA的基因首先在细胞核内经由RNA聚合酶Ⅱ转录产生长度在几百至几万nt的初始miRNA（primary RNA，pri-RNA）。部分研究也发现一些miRNA的转录与RNA聚合酶Ⅲ有关。pri-miRNA被一种称为微处理器（microprocessor）的多蛋白质复合物剪切为长度在60~100nt间，具有发夹结构的单链前体miRNA（pre-miRNA）。并通过转运蛋白Exportin-5及其Ran-GTP辅因子，将pre-miRNA转运至细胞质中。Exprotin-5与Ran-GTP形成的复合物对pre-miRNA具有高亲和力，能够自miRNA在细胞核内产生开始一直到被第二次裂解的过程中，对进行miRNA保护。在细胞质中，pre-miRNA经过Dicer酶加工，形成长度在19~24nt的miRNA-miRNA*双链。随后，细胞中

的TRBP以及PACT进行成熟miRNA链的选择（miRNA*链被特异的降解掉），并募集Argonaute（AGO）蛋白与Dicer形成三聚体复合物，进而启动RNA诱导的沉默复合物（RNA-induced silencing complex，RISC）的装配。在哺乳动物中，miRNA通过引导RISC到靶mRNA的结合位点，使得具有内切酶活性的AGO蛋白能够对靶向的mRNA进行降解。其他的miRNA能够与其特定靶基因的3' UTR部分匹配。这种不完全的碱基配对导致靶mRNA的翻译抑制或者使其脱腺苷化，进而导致靶mRNA的不稳定（图10-1）。

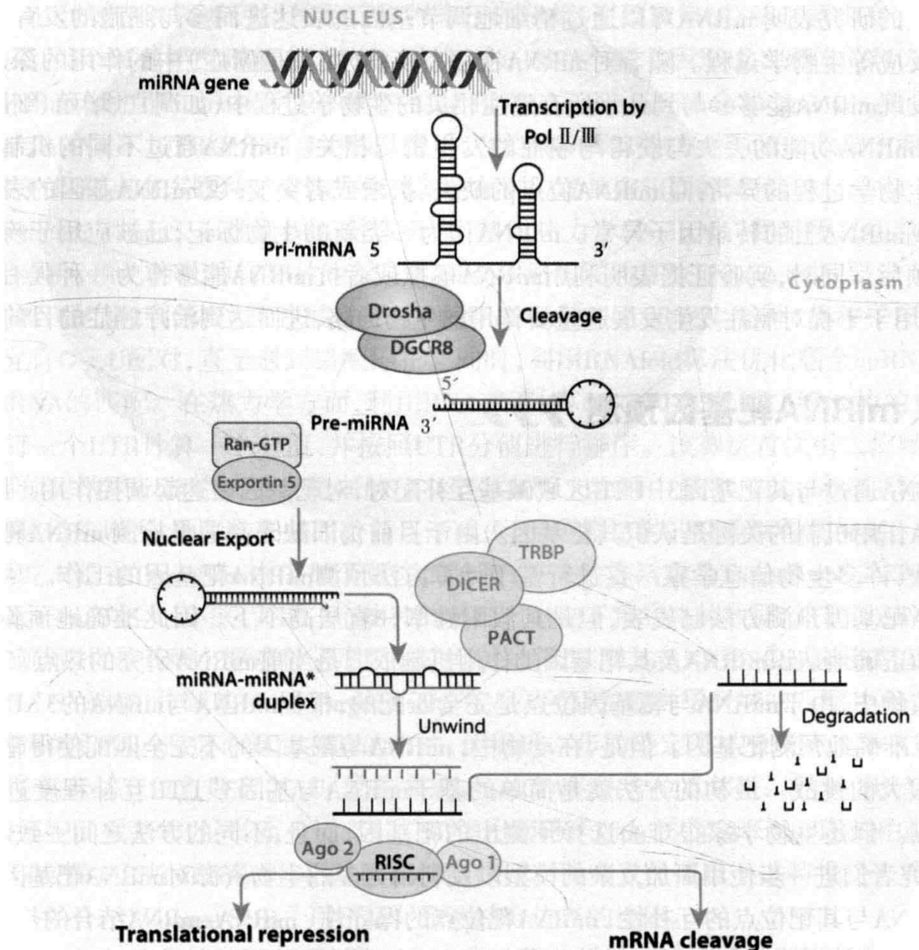


图10-1 miRNA的生物起源

成熟miRNA主要通过抑制和降解两种方式调节靶基因的表达。两种方式的选择取决于miRNA与靶mRNA间的互补程度，即“种子区域”（通常指miRNA 5' 端2-8的核苷酸序列）与靶mRNA 3' 端的互补性。如果两者完全互补则miRNA使mRNA降解，否则对mRNA进行翻译抑制。根据与靶基因的结合方式不同，miRNA大致分为三类：第一类以线虫中的Lin-4为代表，该类miRNA与靶基因以不完全互补的方式结合，抑制mRNA的翻译但不影响其稳定性。目前所发现的大部分miRNA属于此类；第二类以拟南芥中的miR-171为代表，该类miRNA与

靶基因以完全互补的方式结合，其作用方式和功能与siRNA非常相似——直接降解靶mRNA；第三类以Let-7为代表，该类miRNA可以通过以上两种结合方式作用于靶基因。如在果蝇中的Let-7直接介导RISC降解其靶mRNA，而在线虫中的Let-7则与靶mRNA 3' UTR以不完全配对的方式结合，进而抑制其靶mRNA的翻译。在哺乳动物细胞中，蛋白质组实验表明单个miRNA能够直接抑制上百种蛋白质的产生。而且，这些蛋白的抑制绝大多数是由于mRNA表达水平的下调以及翻译抑制所引起的。然而，这种经由miRNA诱导的抑制作用却并不强。有趣的是，在某些条件下，miRNA能够上调其靶mRNA的翻译，或者甚至能够直接对其靶基因的转录进行干预。目前，对于这种罕见的调控模式所知甚少。

大量的研究表明miRNA可以通过精细地调节基因的表达进而参与细胞的发育、分化以及应激反应等生物学过程。随着对miRNA在复杂疾病（尤其是癌症）中的作用的深入研究，研究者发现miRNA能够参与到几乎所有癌症相关的生物学过程中（如凋亡、增殖、细胞周期、转移）。miRNA功能的丢失与获得与癌症的发生密切相关。miRNA通过不同的机制引起癌症相关生物学过程的异常：①miRNA位点的缺失、扩增或者突变；②miRNA基因的表观沉默；③结合到miRNA上的转录因子异常。miRNA作为一类新的生物标记，已被应用于疾病的诊断以及预后。同时，实验证据表明利用miRNA模拟或者抗miRNA能够作为一种强有力的治疗手段，用于干扰对癌症发生发展起重要作用的生物通路，进而达到治疗癌症的目的。

二、miRNA靶基因预测

miRNA通过与其靶基因3' UTR区域碱基互补配对，对靶基因表达起调控作用。因此，认识miRNA作用机制的关键是认识其靶基因。由于目前仍旧缺乏高通量检测miRNA靶基因的实验方法，许多生物信息学家一直进行基于计算方法预测miRNA靶基因的工作。尽管大量的miRNA靶基因预测方法已发表，但是其假阳性率一直居高不下。因此准确地预测miRNA靶基因和正确地认识miRNA及其靶基因的作用机制依旧是当前miRNA研究的热点。

在植物中，由于miRNA与靶基因位点是完全匹配的，根据miRNA与mRNA的3' UTR序列配对可以准确地预测靶基因。但是，在动物中，miRNA与靶基因的不完全匹配使得靶基因预测面临很大的挑战。最初的方法就是简单的基于miRNA与基因3' UTR互补程度进行靶基因的预测。但是生物学家很难去选择预测出的靶基因，而且，不同的方法之间一致性很差。随后，研究者们进一步使用更加复杂的模型以及利用更多的生物资源对miRNA靶基因进行预测（如miRNA与其靶位点的互补性、miRNA靶位点的保守性、miRNA-mRNA结合的热稳定性、miRNA靶位点处不应有复杂二级结构以及miRNA 5' 端与靶基因的结合能力应强于3' 端）。

目前，主要的生物信息学算法包括miRanda、TargetScan、PicTar等。基于序列的miRNA靶基因预测算法虽然各不相同，但通常遵循以下几个原则：①miRNA与靶基因互补性。miRNA的靶点通常分为三类，为5' 端主导型、5' 端种子主导型和3' 端互补型。“种子区域”是指从miRNA 序列5' 端第2个核苷酸起向3' 端延伸连续7个核苷酸（2-8nt）。5' 端主导型是指miRNA的5' 端和3' 端都具有较好的碱基互补配对；5' 端主导种子型是指miRNA的3' 端没有发生较好的碱基互补配对，但miRNA的5' 端至少有连续的7个碱基与mRNA的3' UTR完全互补；3' 端互补型是指miRNA序列3' 端多个碱基与靶基因发生互补配对，但种子区域匹配不充分。②靶点在多物种间的序列保守性。③miRNA与mRNA形成双链结构的热力学稳定

性。④靶基因二级结构和靶点外的序列对靶基因预测的影响。

miRanda: miRanda是2003年提出的一个miRNA靶基因预测软件。它依据miRNA与mRNA序列匹配程度、miRNA与mRNA二级结构热稳定性以及靶位点在物种间的保守性对mRNA的3' UTR进行分析。miRanda首先采用类似于Smith-Waterman的算法对miRNA和mRNA的3' UTR序列进行碱基互补分析，构建打分矩阵：如果正确互补配对（如G：C和A：U），配对分数为+5；错配罚分为-3，起始空位罚分为-8，延伸空位罚分为-2；如果是G：U配对，配对分数为+2。为体现miRNA 的5' 端和3' 端在与靶基因结合过程中作用的不均一性，5' 端的前11个碱基的互补分值需乘以一个尺度参数。碱基互补遵循4个规则：miRNA第2~4位碱基必须和靶基因完全匹配；第3~12位碱基和靶基因错配数目不得多于5个；第9至倒数第6位碱基至少有一个错配；miRNA的最后5个碱基错配不能多于2个。其次，在miRNA与靶基因形成二聚体的热力学稳定性方面，miRanda利用Vienna软件包中的RNAlib（RNAsecondary structure programming library）计算miRNA与mRNA 3' UTR结合的自由能。最后，在物种间保守性方面，miRanda要求靶点在多物种间保守，即靶点在多物种3' UTR序列比对中相同位置具有相同的碱基。

TargetScan: TargetScan是由Lewis等人开发，基于热力学的miRNA-靶基因二级结构特征和保守性分析，预测哺乳动物物种间保守的miRNA靶基因的算法。TargetScan要求"种子匹配"，即miRNA的第2到第8位核苷酸种子序列和mRNA的3' UTR完全互补，从种子序列向两端延伸，允许G：U配对，直至遇到错配停止。同时，利用RNAfold算法优化剩余miRNA 3' 端区域与mRNA的匹配。在热力学方面，利用RNAeval计算miRNA-靶基因二级结构的自由能。最终，对每一个UTR计算一个分值，并按照UTR分值进行排序。该算法首次引入信噪比来评价靶基因预测结果。该算法要求靶向的UTR至少在两个物种中保守。TargetScan算法发现，随着物种数目的增多，预测的靶基因数目逐渐减少，但预测结果的准确率得到提高。2005年，同一组研究人员在TargetScan中添加了更多的物种，改进的算法称为TargetScanS，与TargetScan相比，TargetScanS在人、小鼠、大鼠三个物种的基础上增加了狗和鸡的数据，并重新定义了种子序列（第2到第7位核苷酸），要求种子序列完全互补的情况下，miRNA第8位碱基和靶基因互补或者miRNA 5' 端第1位碱基是A。TargetScan研究人员随后发现种子区域的匹配并不一定会引起其靶基因的抑制。通过计算和实验的方法，他们进一步的确定了结合位点上下文相关的5个特征：①靠近结合位点处富含AU；②与共表达miRNA的结合位点邻近；③与miRNA第13~16个核苷酸匹配的残基邻近；④至少远离3' UTR终止密码子15nt；⑤远离长的UTR的中心。这些特征能够有效的反映miRNA对其靶基因的抑制作用。研究人员通过上述特征进一步改善了TargetScan，并引入"Context Score"用于量化预测结合位点的性能。

PicTar: PicTar开发于2005年，是第一个结合机器学习对miRNA靶基因进行预测的方法。该算法兼顾了靶基因预测算法的基本思想，同时引入了机器学习方法提取特征参数，从统计的角度反映miRNA和靶基因相互作用的显著性。PicTar算法的前提假设是miRNA的不同组合在不同细胞系中可能协同地调控细胞特异基因的表达。PicTar以多重序列比对的3' UTR和共表达的成熟miRNA作为输入，用nuclMAP预测UTR序列上所有可能的miRNA靶位点，检测其miRNA和靶基因二聚体是否符合结合能标准，然后过滤掉没有足够靶位点的3' UTR，并利用隐马尔科夫模型最大似然法对每个UTR打分，最后进行排序。PicTar在miRNA与靶基因序列匹配时，把种子序列分为"完全匹配的种子序列"和"不完全匹配的种子序列"，后者在满足结合能标准前提下允许种子序列出现错配，但不允许G-U配对。同年提出的TargetBoost

算法，是把遗传算法应用到靶基因预测中。从miRNA及其靶基因相互作用特征提炼出加权的模序作为输入参数，对于不同的miRNA进行分类，然后返回不同miRNA和靶基因相互作用的概率作为靶基因的得分。该算法不依赖于靶基因物种间的保守性，且通过提高特征参数的质量就可大幅提高预测准确率。2006年，miTarget 把支持向量机算法融入到靶基因预测中。算法用径向基核函数作为相似指标，把支持向量机特征分为三类（miRNA和靶基因二级结构特征、热力学特征及miRNA和靶基因相互作用的碱基位置特征），接着评估特征参数，最后对预测的靶基因进行打分。

GenMiR++: miRNA在转录后调控水平起了很大的作用，与其靶序列进行匹配，抑制mRNA翻译起始或降解mRNA。因此，miRNA对基因mRNA水平的调控具有很大的贡献。结合miRNA和mRNA表达谱为预测miRNA靶基因提供了新的思路。由于miRNA下调其靶mRNA表达水平，miRNA和它的靶点在表达谱上呈逆向关系。2007年，Huang等人检测88个组织的miRNA和mRNA表达数据，并基于这种逆向关系，利用miRNA-mRNA表达谱构建miRNA-靶基因调控网络，开发了基于贝叶斯方法的靶基因预测算法GenMiR++。他们发现了104个人类miRNA的高精度的靶基因，并通过实验证实了预测的let-7b靶基因。研究结果表明，与基于序列的方法相比，利用相同样本中同时检测miRNA和mRNA的表达谱可以更准确的预测miRNA靶基因。

在当前的miRNA靶基因预测研究中，研究人员逐渐认识到许多新的miRNA-靶基因结合特征。例如，研究发现，miRNA与靶基因结合的过程中，mRNA 的3' UTR二级结构起着重要作用——miRNA靶点几乎都落入3' UTR的二级结构不稳定区域内，然而提高靶点附近序列二级结构的稳定性能够大大降低miRNA对靶基因的作用。已有实验表明，靶点外的序列也对miRNA调节靶基因起到重要作用。靶点后的一段序列对miRNA与靶基因的识别起着重要的作用，对该段序列突变后miRNA对靶基因的调控作用明显减弱，而将该段序列完全删除后miRNA对靶基因的调控作用完全消失。这些新颖的特征也被逐渐加入到miRNA靶预测算法中，进一步提高算法的精确性。同时，研究人员也逐渐意识到，单一依靠序列信息或者表达信息难以继续提高miRNA靶基因预测效能。因此，整合不同层面的数据信息（如功能信息、蛋白质互作信息、表达信息、序列信息等）以及目前实验已证实的miRNA靶基因资源能够进一步提高miRNA靶基因预测的精确性。此外，最近出现的基于深度测序的miRNA靶基因检测方法也为miRNA靶预测带来了新的希望。这些研究将对揭示miRNA功能、了解miRNA诱导疾病发生的机制以及将miRNA用于癌症治疗等关键问题起到重要作用。

三、miRNA数据资源

(一) miRBase

miRBase是一个主要用于存储miRNA序列及其相关注释信息的在线数据库（网址: http: //www.mirbase.org/ ）。当前的miRBase版本（miRBase 18）包含超过18000个发夹pre-miRNA，代表了来自168个物种的21 000个成熟miRNA。它是一个集miRNA序列、注释信息以及预测的靶基因数据为一体的数据库，是目前存储miRNA信息最主要的公共数据库之一。该数据库主要包括三部分内容，即miRBase Registry、miRBase Sequence以及miRBase Targets，

其主要目的：①保持所有miRNA的命名规则一致，并且为新发现的miRNA进行命名；②存储所有已发现的miRNA序列，提供便捷的网上搜索服务以及所有miRNA数据的批量下载；③提供miRNA相关的注释信息（如功能数据、基因组定位、相关参考文献）；④为用户提供来自不同靶预测算法的靶标信息的外部链接。

图10-2所示为miRBase的主界面。主界面包含四部分：①miRBase数据库的更新信息，用户可以很方便地知道miRBase的下个版本的发布日期；②miRBase的版本号（点击其版本号可以看到其当前版本的整体数据统计，及其与前一版本的差异）、便捷的搜索栏、批量下载的入口；③miRBase数据库所提供的基本信息描述；④miRBase数据库的相关参考文献。通过在搜索栏中输入特定miRNA的名字，点击"GO"则可以迅速查询到该miRNA相关的基本信息。该信息页面提供了大量关于该miRNA的基本信息，包括：名字、茎环结构、深度测序相关信息、基因组定位信息、同簇miRNA及其外部数据的相关链接。同时，该页面还包含了该pre-miRNA所产生的成熟miRNA信息，包括成熟miRNA名字、序列及其相关预测的靶基因。最后，该页面提供了与该pre-miRNA相关的所有参考文献。

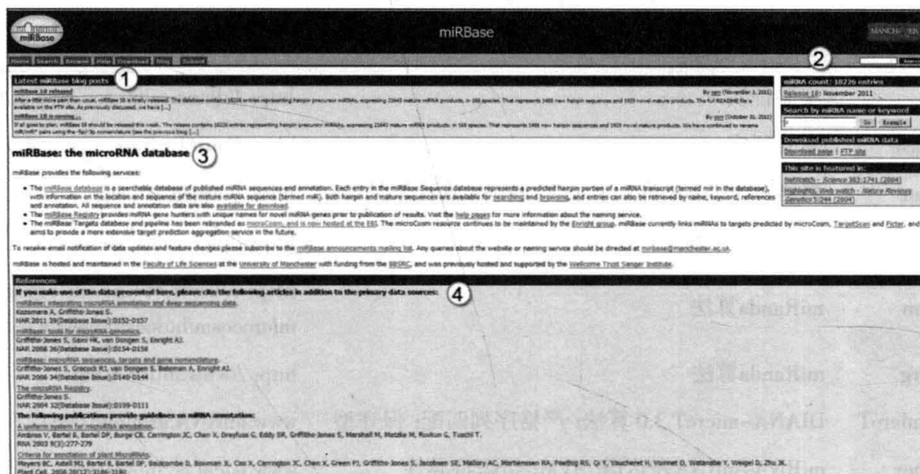


图10-2　miRBase数据库主界面

以hsa-let-7a-1为例，图10-3显示该pre-miRNA所有信息。

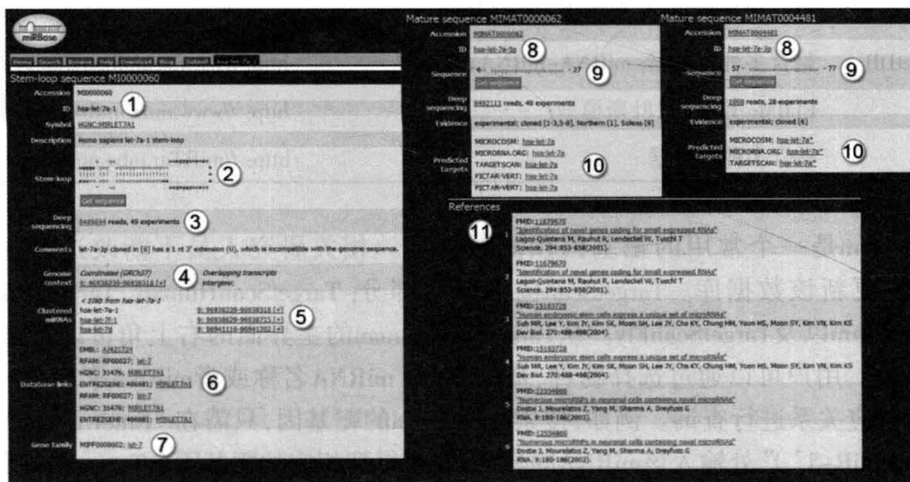


图10-3　miRBase数据库查询结果解析图

①miRNA在miRBase数据库中的名字为hsa-let-7a-1；②hsa-let-7a-1的茎环结构；③来自49个实验的8489694 reads，点击可以查询在49个实验中与该pre-miRNA相关的read数目的具体统计信息；④Hsa-let-7a-1位于9号染色体的正链，具体区间为96938239-96938318；⑤与hsa-let-7a-1同一簇的其他miRNA及它们的基因组定位信息；⑥外部数据库的相关链接（如Rfam数据库）；⑦hsa-let-7a-1属于let-7a家族；⑧hsa-let-7a-1的成熟miRNA（hsa-let-7a-5p以及hsa-let-7a-3p）；⑨hsa-let-7a-5p以及hsa-let-7a-3p的序列分别为ugagguaguagguuguauaguu和cuauacaaucuacugucuuuc；⑩不同靶预测算法预测的靶基因（如MICROCOSM、MIRNA.ORG、TARGETSCAN、PICTAR）；⑪与该miRNA相关的参考文献。

（二）miRNA靶基因数据库

目前，研究人员开发了大量的miRNA靶基因数据库，包括被实验验证的靶基因数据库，基于某种预测算法得到的靶基因数据库以及整合多种预测算法结果的数据库（表10-1）。

表10-1 miRNA靶基因数据库

名称	特点	网址
TarBase	实验证实	http://diana.cslab.ece.ntua.gr/tarbase/
miRTarBase	实验证实	http://miRTarBase.mbc.nctu.edu.tw/
TargetScan	严格种子匹配；位点上下文；保守（或者非保守）	http://targetscan.org
PicTar	严格种子匹配	http://pictar.mdc-berlin.de
MicroCosm	miRanda算法	http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/
MiRNA.org	miRanda算法	http://www.miRNA.org
DIANA-microT	DIANA-microT 3.0 算法；严格序列匹配；保守型	www.miRNA.gr/microT-v4.
TargetMiner	miRNA-mRNA表达谱；SVM；组织特异性	www.isical.ac.in/~bioinfo_miu
RepTar	不依赖结合位点的保守型；不限于种子区域的匹配	http://reptar.ekmd.huji.ac.il
miRror	整合多种靶数据；集中于miRNA协同	http://www.proto.cs.huji.ac.il/mirror
ExprTargetDB	整合多种靶数据；miRNA-mRNA表达谱	http://www.scandb.org/apps/miRNA/
MirZ	ElMMo 算法；贝叶斯模型	http://www.mirz.unibas.ch
miRTar	整合多种靶数据	http://miRTar.mbc.nctu.edu.tw/

TargetScan是一个常用的靶基因预测数据库。相关研究人员不断改进TargetScan算法，并及时更新该数据库。TargetScan包含四个部分：TargetScanHuman、TargetScanMouse、TargetScanWorm以及TargetScanFly。从TargetScanHuman的主界面的右上角可以点击进入其他三个部分。用户可以通过选择物种、基因名称、miRNA名称或者miRNA家族对miRNA与靶基因的对应关系进行查询。例如，搜索hsa-let-7a的靶基因，只需在“Enter a miRNA name（e.g.‘mmu-miR-1’）”处输入该miRNA的名字，即可得到相应的靶基因。TargetScan搜索结果提供了丰富的信息，包含不同结合位点的不同种子匹配类型、结合位点的保守型与“context

score” 等。这些信息能够从不同的角度反映该靶基因预测结果的准确性。需要注意的是，该结果只提供了具有保守靶位点的靶基因。TargetScan也提供了不考虑靶位点保守性的靶基因集合。通过点击 “[View top predicted targets, irrespective of site conservation]”,即可看到大量的不考虑靶点保守性的靶基因集合。此外,TargetScan提供所有数据的批量下载。

第二节

miRNA转录组

Section 2 miRNA Transcriptome

一、miRNA表达谱识别癌症相关miRNA

miRNA作为一类重要的基因调控因子，通过与靶基因3' UTR结合，广泛参与各种生物学过程，因此检测疾病发展过程中不同时期或不同状态下miRNA的表达并识别异常的miRNA对疾病诊断或预后有很大的帮助。近年来，很多癌症研究表明miRNA表达谱能够有效地分类癌症，并且miRNA表达变化与癌症的发生、发展及转移密切相关。因此，基于miRNA表达谱来挖掘人类疾病相关的miRNA是现在癌症研究重点之一。

(一) miRNA表达谱种类

人们发现miRNA在癌症发生发展过程中起着重要作用，检测并分析miRNA表达谱成为研究miRNA功能的一个重要的部分。随着对miRNA序列结构了解的深入，越来越多的技术被用于检测miRNA的表达。除了传统的芯片检测技术之外，许多新的miRNA表达检测技术应运而生。目前应用于检测miRNA表达水平的其他生物学方法还包括克隆、Northern印迹、定量实时PCR（quantitative real time polymerase chain reaction，qRT-PCR）、原位杂交（in situ hybridization，ISH）和新一代高通量测序技术（next-generation sequencing，NGS）等。这些方法已经成功应用于检测miRNA的表达研究。根据实验检测技术，常用的miRNA表达谱可以分为定量实时PCR（qRT-PCR）miRNA表达谱、芯片杂交产生的miRNA表达谱和新一代高通量检测的miRNA表达谱。检测miRNA表达谱平台主要包括Agilent、Exiqon、Illumina、Ambion、Combimatrix、Invitrogen等。基于以上方法和平台检测的miRNA表达谱数据已经陆续被提交到GEO或ArrayExpress等公共数据库中。大量利用miRNA表达谱的研究已经完成并被陆续发表。这些研究涉及疾病诊断、预后及疾病发生发展相关的miRNA标记等方面，其中癌症相关的miRNA表达研究占各种疾病研究的主要部分。已有大量癌症miRNA标记被证实在癌症发生、发展或转移过程中起到重要作用。随着mRNA表达谱检测技术的不断成熟，日益累积的miRNA表达谱数据为今后进一步研究癌症致病机制提供了重要数据来源。

(二) miRNA表达谱检测技术的差异

尽管利用DNA芯片技术来检测miRNA的表达越来越受欢迎，但是这些技术准确性并没有被充分地证实。早期研究表明，不同芯片技术检测同样本mRNA基因表达的可重复率较

低，相关性较差。这种不一致性部分可能是由早期平台的不完整或不正确的注释，技术之间探针的不正确匹配或数据标准化的差异性导致的。随着技术的发展，更精确的注释平台、更合理的探针匹配以及进一步优化的数学模型和过滤技术使得芯片检测技术内部或技术之间有了相对高的一致性。

检测基因表达的技术虽然可以用来检测miRNA的表达，但是也面临着很大的挑战。就芯片技术而言，miRNA序列长度较短直接限制了探针的设计，可能整个miRNA被看做一个探针。同时，由于miRNA家族的存在，家族成员之间序列相似性很高，导致设计的探针呈现冗余性，不能够很好检测不同成员之间差异的表达模式。由于新的miRNA序列长度较短且不同平台间的探针设计和实验协议有着显著的不同，研究不同技术平台检测miRNA表达的一致性是非常有必要的。目前，将TaqMan PCR实验检测的miRNA表达谱做为金标准，通过与TaqMan PCR 实验检测的miRNA表达谱相比较来分析四个miRNA芯片技术之间的准确性和可重复性。在本文中，应用两个商业常用的小鼠参考RNA来创建两个参考样本池，这两个参考样本池可以保证样本间miRNA丰度的最大差异。参考RNA样本池1来自小鼠胚胎睾丸、卵巢和胚胎，参考RNA样本池2来自小鼠胚胎肝、心、肺三个组织。把这两参考RNA池分成四等份，分别利用4个全基因组范围内芯片技术（Agilent、Exiqon、Invitrogen NCode and LC Sciences）进行杂交产生miRNA表达谱。利用不同芯片技术检测miRNA表达谱，对技术内和技术间进行比较，同时和同样本TaqMan PCR 实验检测miRNA表达谱进行比较。利用斯皮尔曼相关系数作为衡量准确性和重复性的指标。通过与背景信号相比，54个 miRNA表达同时被所有的平台检测到。平台内miRNA表达谱的重复性很高，斯皮尔曼相关系数大于0.9。不同平台检测的miRNA表达谱之间仍然有很高的相关系数（变化范围0.663~0.949）。同时，与TaqMan PCR 实验检测的miRNA表达谱比较，这些芯片技术检测的表达谱和TaqMan PCR实验检测的miRNA表达谱具有很高的斯皮尔曼相关系数和一致相关系数。这些结果显示miRNA芯片平台可以产生高度重复数据并且适用于研究miRNA的差异表达。

二、miRNA表达谱分类人类癌症

（一）利用表达谱数据识别癌症相关miRNA

挖掘疾病相关的miRNA已经成为现今非编码RNA研究领域内重点目标之一。大量研究发现miRNA在复杂疾病发生过程中起着非常重要的作用，很多miRNA芯片技术应运而生，并应用于识别复杂疾病特别是癌症相关miRNA。利用miRNA表达谱数据识别癌症相关miRNA一般分为三个步骤。第一步，表达数据获取，即从公共表达数据库（GEO、ArrayExpress等）中查询并下载带有癌症与正常样本的miRNA表达数据。第二步，表达数据预处理，即对所获取的miRNA表达谱数据进行标准化以便于后续分析。然而目前miRNA表达谱数据预处理面临的重要问题是缺乏统一的标准化方法。随着各种miRNA表达检测技术的发展尤其是高通量检测技术的发展，以往mRNA表达谱数据的标准化的方法无法有效地移植到miRNA表达谱数据应用中。第三步，识别癌症相关miRNA，即利用miRNA表达谱寻找癌症发生、发展或转移过程中异常表达（包括上调和下调）的miRNA。寻找异常表达miRNA过程中经常利用到的生物统计学方法有Fold change、*t*检验、SAM、ANOVA等。通过统计学筛选，寻找出癌症样

本和正常样本中差异表达的miRNA，将这些miRNA作为与该种癌症发生密切相关的miRNA。然后在进一步实验中，通过对这些异常miRNA进行敲除或者过表达，对差异表达结果进行生物学证实，从而确定真实癌症相关的miRNA。

（二）利用miRNA表达谱数据分类人类癌症

过去的20年里，在分子水平上的癌症分型研究已经获得巨大成功并被广泛应用于识别癌症相关的生物学标记（biomark）。这些研究表明利用编码蛋白的转录本（mRNA）可以有效地区分各种癌症或癌症的不同亚型，因此这些与癌症相关的转录本可以作为可靠的生物学标记用于癌症致病机制的研究。近几年来，随着生物学界对非编码RNA研究力度的加大，各种非编码小RNA实验检测技术得到快速发展，越来越多小的非编码RNA被发现。这些非编码RNA相应的功能也得到研究和证实。值得注意的是，占这些非编码小RNA研究中比例最大的是对miRNA的研究，大量研究已经证明miRNA的表达异常通常与癌症的发生、发展或转移有密切关系。因此，很多研究已经开始利用miRNA表达谱数据对癌症进行分类，并且将miRNA作为一种新的生物学分子标记用来判断癌症发生、发展或者预后。

2005年的Nature期刊中，Lu等人成功地利用磁珠流式细胞术检测技术系统检测到了涉及多种癌症的334个样本，其中包含了217个人类miRNA的表达水平。Lu等人发现miRNA表达谱中含有大量能够准确反映发育谱系和肿瘤的分化状态的信息，他们观察到与正常样本的表达水平相比miRNA在肿瘤样本中的表达普遍下调。在研究中，Lu等人首次全面证实了利用miRNA对癌症分类具有有效性及可行性。随后，大量的miRNA表达谱研究证实了miRNA作为生物学分子标记的可靠性。在本小节中，我们将探索如何利用miRNA表达数据对癌症进行分类。Lu等人检测的334个样本中包括多种人类组织，包括乳腺、前列腺、胃、结肠和肺等，其中某些组织样本取自癌症患者，例如肺癌、乳腺癌、白血病等患者。从GEO中获取334个样本的原始miRNA表达数据并进行预处理。预处理过程中，基于两套miRNA探针集所包含的控制探针对所有miRNA探针检测值进行标准化，对表达强度偏低的探针进行修正。之后，删除探针集中所有控制探针，并对miRNA探针集检测到的表达值进行以2为底的对数转换。基于miRNA表达谱，利用层次聚类方法对218个样本进行聚类分析。从聚类图中可以看出几乎所有的miRNA表达值在不同的癌症类型中都具有差异。聚类图显示具有共同组织发育起源的样本都被聚到一起。例如，来自结肝、肠、胰腺以及胃部的样本被很好聚在一起，这些不同的组织样本共同起源于胚胎的内胚层；起源于上皮组织或胃肠道组织的样本全部被聚到一起形成一个聚类分支；而造血相关的恶性肿瘤样本明显地分布于另一主要分支上。聚类结果表明miRNA表达谱能够很好区分不同组织起源的样本。为了进一步证实miRNA表达谱能应用于癌症诊断，研究人员选取了68个高分化的癌症样本（代表11种不同的组织类型），利用概率神经网络算法对这些样本的miRNA表达谱数据分别进行训练并产生相应的多类别的分类器。然后，利用训练产生的分类器对17个低分化的肿瘤样本的组织类型进行预测。基于miRNA表达值进行训练的分类器正确分类了17个低分化肿瘤样本中的12个样本。Lu等人的数据中还包括了来自218个样本中的89个组织的mRNA表达谱数据。利用这些mRNA表达谱数据（大约包含16 000个mRNA）对同样的样本进行聚类时，发现具有相同组织起源的样本并没有被聚到一起。同时，利用mRNA基于68个高分化的癌症样本构建表达谱构建神经网络分类器，并对17个低分化肿瘤样本进行检测，结果表明只能正确分类其中

的一个样本。该数据还包含来自73个急性淋巴细胞白血病患者的骨髓样本的miRNA表达水平。经过层次聚类，一个主要分支包含所有5个BCR/ABL阳性样本以及11个TEL/AML1样本中的10个样本；另一个主要分支包含了19个急性T细胞淋巴细胞白血病样本中的13个样本。聚类结果说明即使对于同一组织起源的样本，利用miRNA表达数据进行聚类仍旧能够得到疾病的不同亚型并识别不同亚型中miRNA表达模式。

由于miRNA在癌症分类中具有的重要作用，越来越多的研究利用miRNA表达谱研究同种疾病的不同亚型。Cherie Blenkiron等利用miRNA表达谱分析乳腺癌并识别肿瘤亚型生物标记分子。该研究分析了包含93个原发乳腺癌、33个乳腺癌细胞系和5个正常乳腺样本的miRNA表达谱。通过层次聚类，发现miRNA表达谱能够很好地把乳腺癌细胞系、原发肿瘤样本和正常样本分开。同时聚类分析结果表明ER-和ER+两个乳腺癌亚型在miRNA表达模式上存在显著的不同。为了进一步证实miRNA是否在乳腺癌不同亚型中差异表达，利用单样本预测算法把93原发乳腺癌样本进行亚型分类：luminal A、luminal B、basal-like、HER2+和normal-like。通过识别乳腺癌亚型间差异表达的miRNA，利用差异表达的miRNA对有亚型标签的样本进行有监督聚类。结果表明这些差异表达基因能够很好地将乳腺癌亚型分开。为了证明miRNA表达谱对样本亚型具有预测的潜能，利用检测137个miRNA表达谱的basal-like样本和luminal A样本进行基于模型的判别分析，并对Lu等检测的11乳腺癌样本进行分类。结果表明基于miRNA表达谱可以有效地对乳腺癌亚型进行分类。

这些研究暗示着miRNA表达数据中蕴含着惊人的信息量，不仅能够有效地反映出不同的组织起源和癌症分化状态，而且同mRNA数据相比较，利用miRNA表达谱数据能够更有效地预测出低分化癌症样本的组织类型。总之，miRNA表达谱数据为癌症的诊断提供了潜在的可能性。

(三) miRNA表达谱数据应用于癌症预后

除了利用miRNA表达谱分类癌症，潜在地将miRNA标签应用于癌症诊断之外，很多研究还表明miRNA标签有可能用于人类癌症的预后。这些探索miRNA在肿瘤发展中作用的研究将研究重点放在治疗策略靶向的miRNA或miRNA调控的通路之中，通过研究不同时期或不同阶段癌症中miRNA标签来用于患者的预后。例如，Hu等人发现在人类血清之中存在稳定表达的miRNA，这些miRNA可以作为潜在的疾病标签预测存活。作为实例，Hu等人利用Ⅰ至Ⅲ期的肺癌和鳞状细胞癌患者血清样本进行研究，通过qRT-PCR芯片检测发现30个长期存活的患者血清中的miRNA表达水平与30个短期存活患者相比具有显著差异。通过检测发现四个miRNA标签：miR-486、miR-30d、miR-1和miR-499可以作为非侵蚀性的预测子用来预测非小细胞肺癌患者的存活时间。通过研究182个急性髓样白血病患者样本中miRNA表达，Ramiro等人发现与正常样本相比，疾病样本中很多miRNA差异表达并且这些miRNA的表达与分子异常紧密相关。通过对122个新诊断为急性髓样白血病患者样本的miRNA表达谱进行生存分析发现miR-191和miR-199a两个miRNA与急性髓样白血病患者的预后不良显著相关。

三、miRNA表达谱与mRNA表达谱整合分析

近年来，miRNA作为转录后调控的重要调控因子成为科研的研究热点。很多研究人员

致力于研究miRNA的功能，并开发了很多miRNA靶预测算法。预测结果表明miRNA和靶基因之间存在多对多的关系，也就是靶的多样性和miRNA的协同性。但是这些预测算法并不能明确指出哪些miRNA靶关系在特定的条件下被激活，进而发挥作用。随着实验检测技术的不断发展，产生了大量的miRNA和mRNA表达谱数据。生物信息学研究者通过整合多种基因组数据（miRNA表达谱、mRNA表达谱、miRNA靶关系、基因本体论、蛋白质互作网络）来研究miRNA的功能。整合miRNA表达谱和mRNA表达谱研究疾病将有助于提高研究结果的准确性。

miRNA和mRNA表达谱可以用来衡量某特定条件下miRNA和基因的活性。同时，miRNA和mRNA表达谱提供了不同细胞状态下miRNA、基因以及二者的调控关系在转录水平上的动态性。整合miRNA表达谱和mRNA表达谱，能够更准确地研究miRNA靶向关系，进而明确miRNA在不同状态下的作用。我们可以从两个方面来分析miRNA在不同状态下的功能。一是基于miRNA表达谱和mRNA表达谱预测miRNA靶基因，进而分析miRNA的功能。二是，结合已知基于序列的靶基因预测算法，利用miRNA表达谱和mRNA表达谱识别在特定条件下的miRNA-mRNA调控模块。

（一）预测miRNA靶位点

虽然很多基于序列靶预测算法已经预测出很多miRNA靶基因，只有很少数的miRNA通过实验证实具有特定的功能。准确预测miRNA靶基因不但是研究miRNA功能特征的一个瓶颈，而且是研究由miRNA失调引发的人类疾病的关键。此外，难以正确的识别具有生理活性的miRNA仍是研究miRNA功能特征的阻碍。

众所周知，miRNA通过抑制靶基因的翻译或降解mRNA来调节基因的表达。miRNA的功能失调会导致下游靶基因的表达紊乱。很多miRNA转染或敲除实验使得其靶基因的表达降低或升高，进而证明miRNA与其靶基因呈现靶向关系，并且这种关系是逆向的。最近，Guo等人通过敲除mir-223以及转染mir-1、mir-15证明miRNA主要是通过降解mRNA导致蛋白质的表达水平下降。利用miRNA和mRNA同时检测的表达谱可以准确预测功能miRNA靶点。

Huang等人在序列预测算法的基础上，结合同时检测的88个组织中miRNA和mRNA表达数据，对miRNA靶关系进行进一步筛选，提高了预测精度，并通过实验证实了其预测的靶基因。Gennarino等人基于miRNA与其宿主基因之间具有强共表达关系，利用miRNA宿主基因表达代替miRNA的表达，通过计算miRNA宿主基因与mRNA的逆向共表达关系预测miRNA靶基因，使靶基因预测准确性有所提高。Liu等人基于89个人类组织中的miRNA和mRNA表达谱，计算miRNA-mRNA对的相关性，利用成熟mRNA的功能来推断miRNA的功能。

同时，人们发现miRNA在疾病研究中起着重要作用，因此需要研究miRNA在特定疾病中的功能。利用同步检测疾病相关的miRNA和mRNA表达谱，我们还可以预测特定疾病条件下被激活的miRNA靶向关系。

（二）识别miRNA调控模块

miRNA的出现使基因调控网络变得更为复杂。miRNA作为新的基因调控网络的重要调控子，其功能成为研究热点。研究人员从计算方法和实验方法两个方面来解析miRNA的功能。在研究miRNA之初，人们主要识别miRNA和它们的靶基因，并为此开发了很多预测靶基

因的算法，通过研究它们的靶基因集合的功能来推断miRNA的功能。靶基因预测算法的结果表明，一个miRNA可以靶向成百上千个基因。同时一个基因也可以被数十个miRNA调控。这种miRNA和基因之间多对多的复杂调控关系，即靶基因多样性和miRNA协同性，促使研究人员假设miRNA通过调控它们共同的靶基因组成一个调控模块参与复杂的生物学过程。

尽管miRNA调控机制的研究有了很大的进展，关于miRNA功能的一些基础的问题还是没有弄清楚。比如：在特定的条件下，哪些miRNA表达了？哪些基因表达了？表达的miRNA和基因之间有什么样的关系？ miRNA通过调节哪些靶基因的表达进而对生物学过程调控？累积的实验结果表明并不是单个miRNA导致表型的变化，而是多个miRNA同时靶向细胞过程中的重要组分进而调控生物过程。Mavrakis实验表明miR-9b、miR-20a、miR-26a、miR-9和miR-223 通过协同调控肿瘤抑制基因PTEN、BIM、PHF6、NF1和FBXW7促进T细胞急性淋巴性白血病的发展。

为了了解miRNA在复杂细胞系统中的调控机制，在miRNA和mRNA复杂调控关系中识别出功能模块非常重要。2005年，Yoon等人提出了miRNA-mRNA调控模块的概念，即一组共同参与相同生物学过程的miRNA和其靶基因。他们基于序列匹配程度来识别miRNA调控模块，体现不出特定条件下miRNA靶关系的激活状态。而miRNA和mRNA表达谱能够很好地反映miRNA和mRNA在特定条件下的激活状态，整合miRNA表达谱和mRNA表达谱对于识别特定条件下激活的miRNA-mRNA调控模块有很大的帮助。随着对miRNA的进一步研究和越来越多的同步检测的miRNA表达谱和mRNA表达谱的出现，识别条件特异的高置信的miRNA-mRNA调控模块成为可能。在本文中，我们介绍两种不同整合miRNA表达谱和mRNA表达谱的方法来识别miRNA调控模块。一是直接整合miRNA表达谱和mRNA表达谱，即基于miRNA和mRNA之间逆向共表达的关系。二是间接整合miRNA表达谱和mRNA表达谱，即基于miRNA表达一致和mRNA表达一致性。

直接整合miRNA表达谱和mRNA表达谱的方法要求miRNA表达谱和mRNA表达谱来自同一组样本，这些配对的表达谱能够同时反映miRNA和mRNA在同一种状态下活性。Peng等检测了30个 HCV阳性或阴性人类肝脏活组织样本的miRNA和mRNA表达谱。miRNA主要是通过降解靶基因mRNA水平来行使功能，具有逆向相关的miRNA-mRNA关系对被认为是在HCV条件下激活的。通过计算miRNA和mRNA之间的皮尔森相关系数，寻找逆向相关的miRNA-mRNA关系对。结合miRNA和靶基因在序列水平上的调控关系，构建出特定条件下激活的miRNA-mRNA二部图。人们可以在这个二部图上寻找最大的完全连接的子二部图。

间接整合miRNA表达谱和mRNA表达谱的方法不要求miRNA表达谱和mRNA表达谱来自同一组样本，但是miRNA表达谱和mRNA表达谱应该针对相同的表型。Joung等人开发了一种基于群体的概率学习算法，通过整合miRNA靶基因信息、miRNA表达谱和mRNA表达谱来识别一致miRNA-mRNA调控模块（图10-4）。miRNA表达一致性、mRNA表达一致性和miRNA与mRNA在序列上的绑定程度是识别miRNA-mRNA调控模块的三个组成部分。计算miRNA和mRNA各自皮尔森相关系数和 miRNA与mRNA在序列上的绑定程度，把这三个参数输入到遗传算法中，通过迭代，使目标函数达到最优化，进而获得miRNA-mRNA调控模块（图10-5）。

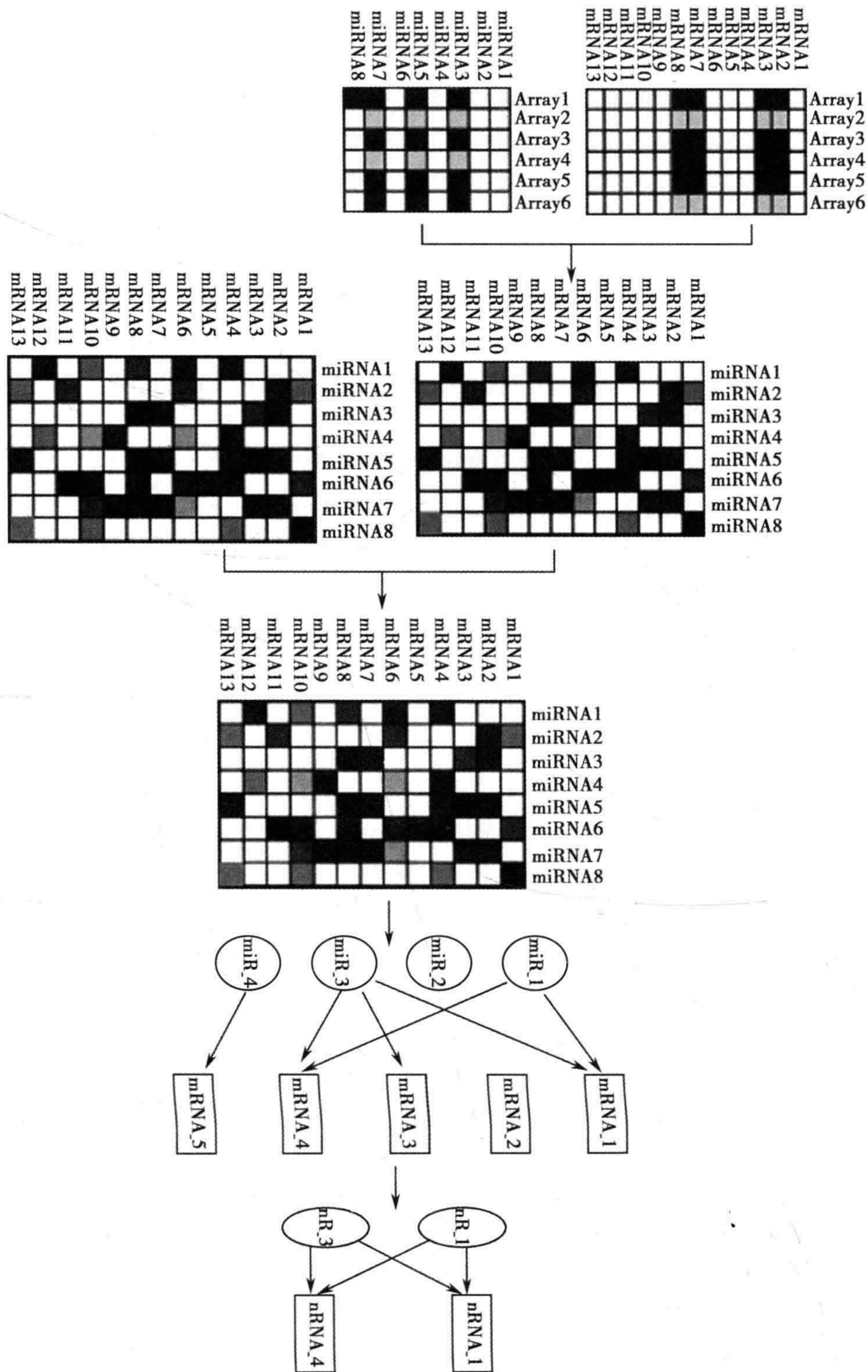


图10-4 基于直接整合miRNA和mRNA表达谱识别miRNA-mRNA调控模块

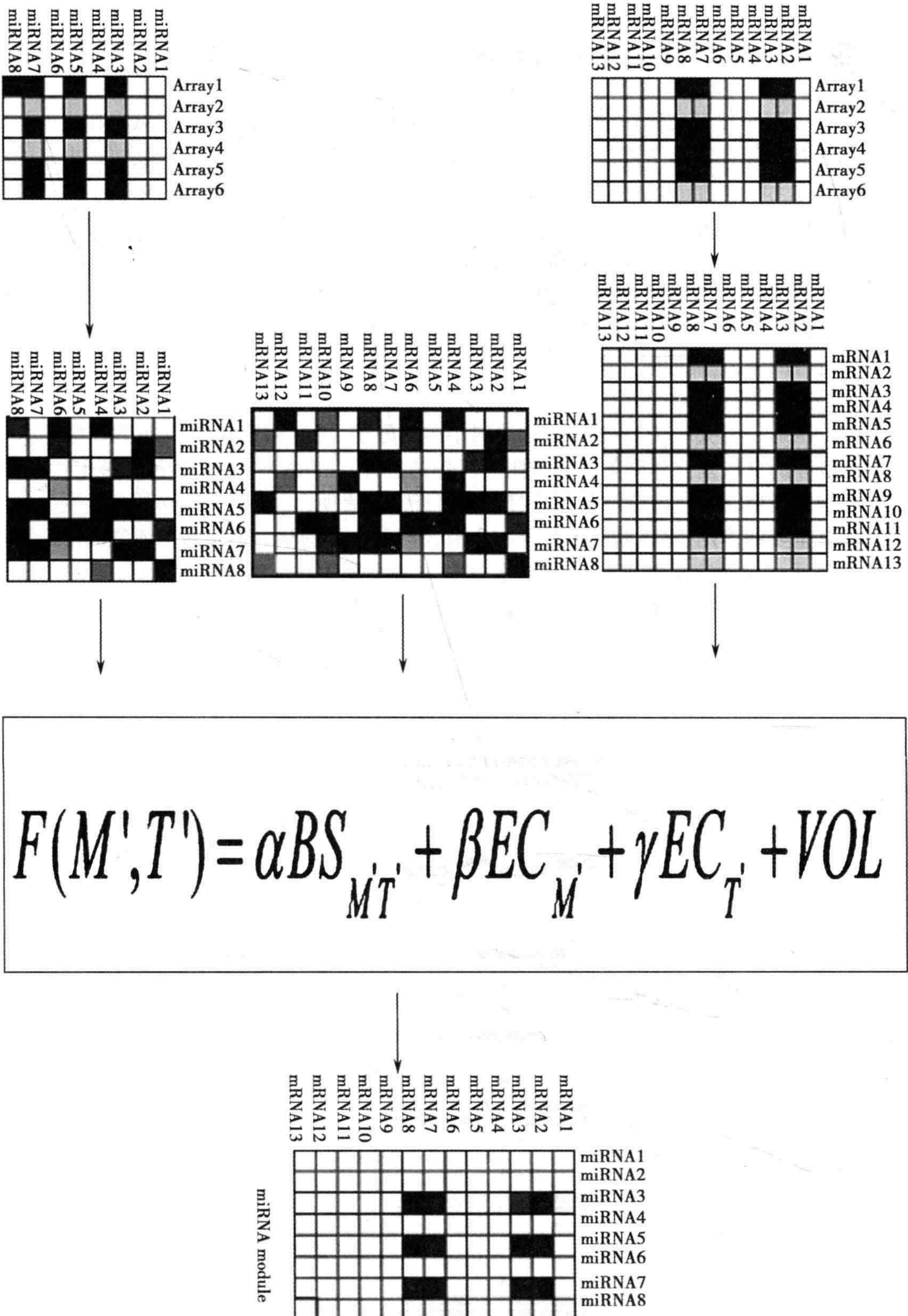


图10-5 基于间接整合miRNA和mRNA表达谱识别miRNA-mRNA调控模块

四、新一代测序检测miRNA转录组

新一代测序技术又称作深度测序技术，主要特点是测序通量高、测序时间和成本与第一代测序技术相比显著下降。转录组是指特定细胞在某一功能状态下所能转录出来的所有RNA的总和，包括mRNA和非编码RNA（non-coding RNA）。非编码RNA又包括：tRNA、rRNA、miRNA、piRNA和long ncRNA等。RNA-Seq 利用高通量测序技术对组织或细胞中所有RNA（即是整个转录组）反转录而成的cDNA 文库进行测序，通过统计相关读段（read）数计算出不同RNA 的表达量，发现新的转录本；如果有基因组参考序列，可以把转录本映射回基因组，确定转录本位置、剪切情况等更为全面的遗传信息。由于RNA-seq是对细胞的整个转录组进行测序，它能同时检测映射的转录区域和基因表达，动态的量化整个转录组的表达水平，区分不同的转录本亚型。

miRNA是一类大小为21~23nt的非编码小RNA分子，通过和靶基因3' 非翻译区结合引导RNA诱导的沉默复合体降解其靶或阻碍其靶的翻译。miRNA存在于各种真核生物中，广泛参与细胞增殖、凋亡、代谢及分化等过程。最近研究表明，miRNA在疾病的发生发展过程中也具有重要的作用，在诊断和治疗疾病上有光明的应用前景。但是目前研究miRNA的主要方法是通过定时定量的PCR进行检测，这些方法主要关注miRNA的表达，并局限于研究那些序列信息和二级茎环结构信息已知的miRNA，无法寻找和发现新的miRNA分子。现在已有专门用于miRNA组的测序技术——miRNA-seq，它能够直接对样本中指定大小的所有miRNA分子进行高通量测序，在无需任何参考序列的条件下研究miRNA的表达谱，并在此基础上鉴定新的miRNA分子，从而进行更加深入的分析（图10-6）。

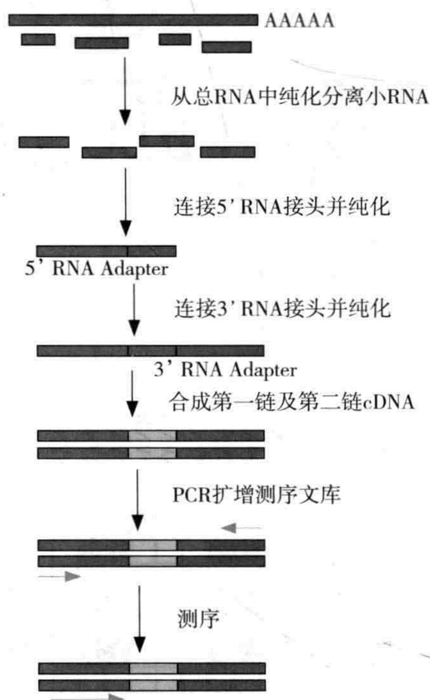


图10-6 miRNA-seq的文库构建及测序过程

（一）miRNA-seq流程

miRNA-seq与RNA-seq不同之处在于文库制备的过程中对于样本的处理，miRNA-seq在文库制备中首先从总的RNA分子中提取长度为21~23个碱基左右的小RNA；然后对提取的RNA 5’端连接接头并纯化，再对3’端连接接头并纯化；然后用随机引物和反转录酶从RNA合成cDNA片段，然后利用凝胶电泳实验对样本进行纯化；cDNA文库制备完成后就可以进行测序了。这样测序得到的将是全部的miRNA转录本，研究人员可以对miRNA-seq数据进行处理分析。miRNA-seq可以一次性获得数百万条miRNA序列，能够快速鉴定出不同组织、不同发育阶段、不同疾病状态下已知和未知的miRNA及其表达差异，为研究miRNA对细胞进程的作用及其生物学影响提供了有力工具。

（二）miRNA转录组分析

1. 数据处理及分析　如图10-7，miRNA-seq的数据处理及分析步骤。对原始数据进行过滤，去除那些可能的测序错误。那些定位到已知miRNA前体序列的read序列经过装配得到miRNA表达数据；那些未被定位到已知miRNA的read序列可以用于发现潜在的新的miRNA；除了miRNA read序列，也可以用于发现其余的small RNA种类、piRNA或者snoRNA。

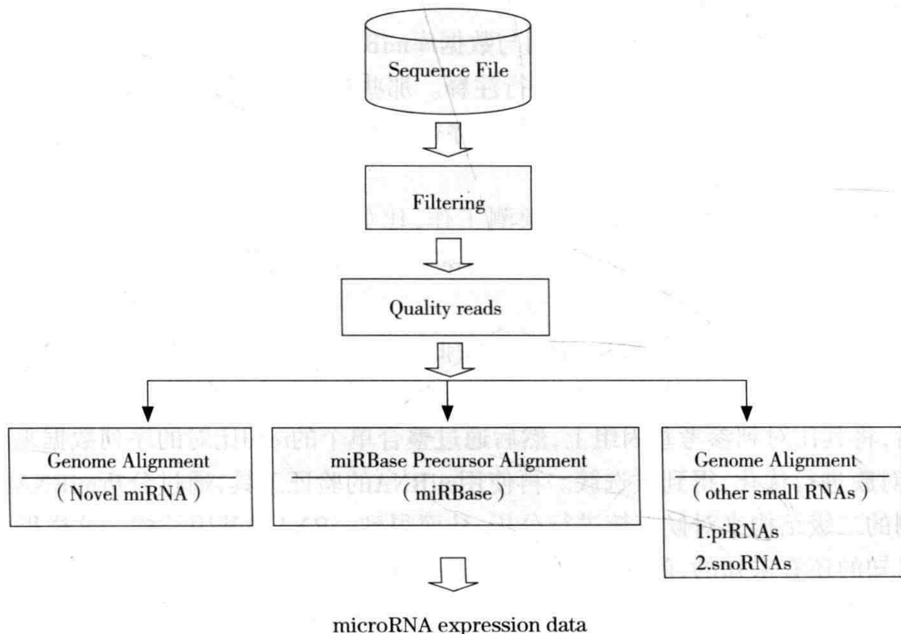


图10-7　miRNA-seq的数据处理及分析步骤

2. 基本数据分析　得到原始数据以后，首先对miRNA-seq的数据进行预处理，例如进行Base-calling，去除污染及接头的序列，过滤质量较差的read等。这样可以得到测定的read的长度、read的数量和其质量。然后将经过预处理的read映射到参考基因组上，可以得到各个read在基因组上的分布。然后将read装配成转录本，根据read在参考基因组上的位置，可以估计出基因的表达水平。read数目与基因真实表达水平成正比，与基因长度成正比，与测序深度正相关。可以用RPKM来衡量基因的表达水平，即每百万读段中来自于某基因每千碱基长

度的读段数。

$$RPKM=\frac{\text{基因区段计数}}{\text{基因长度}\times\text{测序深度}}\times 10^9 \qquad (10\text{-}1)$$

现在已有专门的软件进行这些工作，比如：rseq、DEGseq、Cufflinks等。

3. 高级数据分析　目前常用的生物信息学流程：利用SHRiMP，将miRNA前体发夹序列比对到成熟miRNA和较小的miRNA*，利用得到的编码坐标来确定已知的miRNA。基本处理后，我们得到那些唯一的序列，保留它们的read信息，那些唯一的序列比对到miRBase miRNA前体发夹上来确定成熟miRNA/miRNA*序列。

（三）预测新的miRNA

目前人们已开发出多种算法，来预测miRNA。但是所有方法都利用了二级结构信息，因为发夹结构的存在是miRNA的主要特征。其中许多方法还依靠序列的保守性来区分miRNA候选物和无关的基因组发夹。另一些方法则评估发夹结构与已知miRNA的序列和结构相似度及其热力学稳定性，另一种高效的方法是探索已知miRNA周围的基因组序列，因为许多miRNA都是成簇排布。人和小鼠的许多miRNA就是通过这种方式鉴定出的。当然，计算机预测出来的候选miRNA还需要实验的验证。

我们可以将miRNA-seq得到的序列与数据库miBase、数据库Refseq、rRNA数据库、tRNA数据库进行比对，从而对已知miRNA进行注释。那些在已知数据库中未能找到注释信息的miRNA，则可能是新的miRNA。也可以将测序得到的序列与该物种全基因组序列进行比对分析，通过折叠模型预测新的miRNA。

现在已有专门的软件进行这些预测工作，比如miRAnalyzer，它有三个分析步骤：在miBase数据库中发现有注释的miRNA；再将read定位到转录序列的文库（mRNA、ncRNA）；预测新的miRNA。

利用miRNA-seq得到read数据，首先去除低质量的read、没有3'接头的read或低复杂度的read，选择成熟miRNA长度的read，然后将read处理成那些唯一的序列。获得高质量的唯一序列后，将其比对到参考基因组上，然后通过整合单个的read比对的序列数据来确定序列的簇，再对簇进行优化，得到候选簇。再使用miRNA的验证工具，通过分析miRNA的前体发夹的预测的二级结构来对候选簇进行分析，从而得到miRNA的基因结构。这样得到的基因结构有已知的还有未知的，就可以预测新的miRNA。对于那些未定位到已知miRNA前体的read序列，继续将他们映射到整个基因组上。对于精确定位到基因组上的read利用现有的软件（如Vienna package）折叠成RNA，从而得到一些假定的miRNA发夹结构，然后对这些发夹结构进行过滤，得到具有单个环的发夹结构的假定的成熟的miRNA，这些miRNA作为可能正确的发夹。对他们再进行折叠过滤，就可以得到具有正确茎环结构的新的miRNA。

（四）比较不同miRNA之间的表达差异

利用miRNA-seq数据我们可以得到每个miRNA的表达水平，继而可以比较它们之间的表达差异，也可以根据miRNA所在的簇来分析不同簇的差异表达。可以利用R Bioconductor软件包、DeSeq来进行miRNA差异表达分析。

(五) 发现已知miRNA的新亚型

RNA序列的变异可以导致miRNA的不同的isomiRs。miRNA的isomiRs非常普遍，他们可能是miRNA在生物起源中的修剪或切割所致。通过分析miRNA序列内部变异和3’端变异可以发现miRNA的isomiRs。在比对到miRBase前体发夹后，分析那些没有匹配到miRNA参考基因的3’端变异。那些序列的改变导致该序列没有匹配到前体发夹，根据他们是否与已知的miRNA的编辑过程一致来进行分类。

总之，miRNA-seq方法可以产生关于small RNA的大量数据，即miRNA转录组的数据，很好地刻画miRNA转录组的信息。从这些数据中我们不仅可以得到miRNA的表达信息，发现新的miRNA，预测miRNA的靶基因，检测差异表达的miRNA，还可以得到别的小RNA的信息，例如piRNA、snoRNA。

第三节

miRNA调控网络

Section 3 miRNA Regulatory Network

一、miRNA转录调控网

一个生物体通常由上百种细胞类型组成，这些细胞具有同一套遗传信息，然而它们却在生物体中行使着截然不同的功能。研究者们认为生物体中细胞的状态依赖于基因组的染色质状态。染色质状态能够标记基因组的动态调控模式，也就是说对于不同类型的细胞，它们的基因调控模式不尽相同。细胞或者组织特异的基因调控模式影响基因的表达，进而影响基因编码蛋白质的功能。因此，理解基因的转录调控网络对于揭示细胞发育和分析细胞状态有至关重要的作用。在基因转录调控网络中，尤为重要的两类反式作用因子包括转录因子（transcriptional factor，TF）和miRNA。其中，TF通过特异性地识别靶基因上游5’端特定序列（启动子），与其特异结合进而激活基因的转录。miRNA的种子序列在RNA诱导的沉默复合物作用下特异性识别、绑定靶基因的3’UTR，基于翻译抑制和mRNA降解两种机制在转录后水平调节基因的表达。因此，基因网络中两类重要的调控子（TF和miRNA）在不同层面（转录和转录后水平）调控基因的表达。大量的实验或者计算机靶基因预测算法分析表明，一个调控子（TF或者miRNA）都能够调节多个甚至上百个基因的表达（图10-8），一个基因的表达通常会受到多个调控子的作用。越来越多的证据显示，生物过程中一个基因同时受到两种调控子作用的现象是普遍发生的（图10-9），这说明，TF和miRNA在基因调控网络中存在着互作关系。

目前，研究者们通过基因芯片或者高通量测序技术检测基因的表达，扫描基因组上TF的模体（motif）并分析其富集情况，利用计算机预测或者实验方法识别TF的靶基因，整合这些信息进而构建TF介导的基因转录调控网络。对于miRNA而言，许多靶基因预测算法都能够识别miRNA-mRNA调控关系。同时，在miRNA靶基因预测研究中，研究者发现TF是miRNA

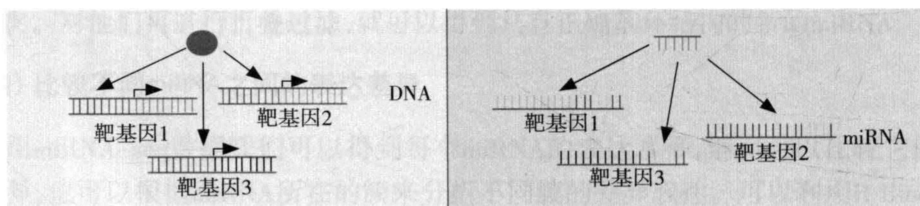


图10-8 TF和miRNA调控多个基因示意图

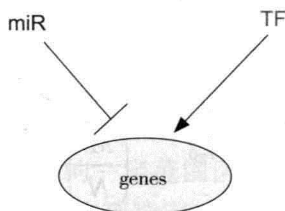


图10-9 两种调控子（TF与miRNA）同时调控一个基因

的一类重要的靶基因，miRNA通过调节TF进而影响下游蛋白质表达，这一证据为基因调控网络中TF与miRNA存在关联提供了支持。随后，越来越多的研究结果都显示在基因表达调控过程中普遍存在TF和miRNA的互作模式，并且基于实验方法证实了调控子互作模式在生物体发育过程中的重要作用。此外，miRNA转染和敲除实验是预测miRNA功能的有效方法，Harnoo等人利用转染癌症相关miRNA得到的基因表达数据，结合TF的靶基因集合，利用Wilcoson检验和K-S检验方法比较TF的靶基因与非靶基因的表达差异情况，进而识别了特定转染的miRNA与相应条件下显著激活的TF的关联，最终构建了与癌症相关的miRNA调控TF网络。该研究结果表明对不同的miRNA进行干扰时，生物体会产生与其相对应的TF应答，进而调节癌症过程中的相关生物学通路。而且，基于双表达谱数据分析，发现TF与miRNA对基因表达的协同调控模式能够为精确分类细胞形态提供可靠依据。因此，识别TF-miRNA互作模式和构建TF-miRNA调控网络对于揭示生物体复杂的生理机制，进而解释复杂疾病的发病机制提供依据。

为了系统性地识别TF-miRNA互作模式，理解基因的转录和转录后调控机制，揭示生物体复杂的调控过程，许多生物信息学研究者已经开始利用各种数据资源来构建TF-miRNA调控网络。

靶基因数据资源：靶基因集合是TF和miRNA功能预测中的一个重要的数据资源，大量研究都利用调控子（TF和miRNA）靶基因的功能来预测分析调控子可能具有的生物学功能，并且许多预测结果已经得到实验证实。那么，靶基因是否可用于预测TF和miRNA的关联？

Shalgi等人通过对TF和miRNA靶基因的重叠现象的研究发现，具有相似靶基因集合的调控子（miRNA对和TF-miRNA对），它们倾向于存在互作关系。因此，基于TF和miRNA的靶基因数据，通过寻找显著共享靶基因的TF和miRNA，能够识别TF-miRNA互作关系。这种互作关系分别在转录和转录后水平控制基因的表达，形成了基因的TF-miRNA共调控网络。

【例10-1】基于共同靶基因构建miRNA转录调控网

1. 数据准备　基于共享靶基因构建miRNA-TF转录调控网络需要两种类型数据：miRNA及其靶基因数据、TF及其靶基因数据。为此，Shalgi等人从TargetScan和PicTar两个数据库中获取保守的miRNA与其靶基因数据，从UCSC中获取TF与其保守的结合位点（TFBS）和miRNA靶基因的序列信息。

2. 网络构建　得到上述数据后，通过寻找miRNA靶基因上是否有TF结合位点来确定该TF的靶基因，再建立miRNA-gene矩阵和TF-gene矩阵，矩阵内有调控关系的元素对取值为1，反之则为0。通过这两个矩阵寻找共享靶基因的miRNA-TF对，再利用超几何检验和随机性检验确定显著共享靶基因的miRNA-TF对。

首先，对每一对miRNA–TF进行超几何检验，计算公式如下：

$$p = 1 - \sum_{i=0}^{x} \frac{\left(\frac{k}{i}\right)\left(\frac{M-K}{N-i}\right)}{\left(\frac{M}{N}\right)} \quad (10\text{–}2)$$

其中，M为含有至少一个TFBS的总的miRNA靶基因个数，N为该对调控子中miRNA靶基因个数，K为该对调控子中TF靶基因个数，x为该对miRNA和TF共享的靶基因个数。计算出p值以后需进行FDR校正，这里取FDR < 0.3。如此我们分别从TargetScan和PicTar中得到111对miRNA–TF和1263对miRNA–TF。

然后，再对上述得到的miRNA–TF对进行随机性检验。我们随机选取具有靶向关系的一对miRNA（i）–gene（1）和一对TF（j）–gene（2）且miRNA（i）与gene（2）、TF（j）与gene（1）没有靶向关系，交换它们的边，即将miRNA（i）–gene（1）和TF（j）–gene（2）的取值由1变为0，将miRNA（i）–gene（2）和一对TF（j）–gene（1）的取值由0变为1，以保证每个miRNA和TF的靶基因个数不变，每个基因对应的miRNA和TF个数不变。该方法为边交换，我们建立1000个随机的miRNA–gene矩阵和TF–gene矩阵，其中每对矩阵都进行了100 000次边交换。我们对这1000个随机得到的矩阵对及原始的矩阵对中的所有miRNA–TF对计算Meet/Min得分：

$$\text{Meet / Min} = \frac{\left|\text{Targets}(i) \cap \text{Targets}(j)\right|}{\min\left(\left|\text{Targets}(i)\right|, \left|\text{Targets}(j)\right|\right)} \quad (10\text{–}3)$$

其中，Targets（j）为第i个miRNA的靶基因集合，Targets（j）为第j个TF的靶基因集合。对于第i个miRNA和第j个TF，随机性检验的p值为1000个随机得到的靶集合对中Meet/Min得分大于原始靶集合对中Meet/Min得分的靶集合对所占的比例，再对所得到的p值进行FDR校正（FDR < 0.3）。

我们发现，超几何检验得到的miRNA–TF对中大部分都通过了随机性检验（TargetScan 92%，PicTar 72%）。经过超几何检验和随机性检验后，我们在TargetScan和PicTar中分别得到104对和916对miRNA–TF对，将这些得到的miRNA–TF对去重以后就可以构建一个简单的miRNA转录调控网络。

Zhou等人利用TargetScan算法和TRANSFAC数据库分别获得miRNA和TF的靶基因集合，同时基于靶基因的表达数据，利用Fisher精确检验和Byesian关联分析算法计算任何一对调控子（TF对、miRNA对和TF–miRNA对）协同调节靶基因的显著性，结果发现大量的TF对和miRNA对共享靶基因，同时，基于调控子间共享靶基因分析，识别了一些TF–miRNA关联对，它们具有显著的靶基因重叠，暗示了这些TF和miRNA在基因调控过程中具有相似的功能。

基因表达数据资源：具有相同或者相似表达模式的基因功能相似，因此，一些研究通过构建共表达网络来预测基因的功能。TF和miRNA通过其靶基因行使功能，所以，基于两类调控子的靶基因表达相似性，识别TF和miRNA关联是可行的。Su等人利用Pearson相关系数计算TF和miRNA靶基因的表达模式相似性，基于相似的靶基因表达模式构建了TF–miRNA模块（module）。结合TF–miRNA关联模块和TF的motif扫描方法，最终构建了基因调控网络，网络包括TF、miRNA、基因以及转录调控和转录后调控关系。

上述的方法基于靶基因或者基于表达数据识别了TF–miRNA关联，然而这些方法并不

能明确TF-miRNA的因果关系。相对于miRNA转录后调节TF的表达而言，如何识别TF调节miRNA的转录要更加复杂和困难一些。原因在于绝大多数miRNA的初始转录本是未被注释的，因此，无法直接从数据库中获得TF在miRNA上的绑定。为了揭示miRNA的转录和识别TF调节的miRNA关系，需要开发有针对性的方法识别miRNA上游的调控元件。最常用的方法是基于motif扫描，在miRNA前体上游的一定范围区间内（如2kb、5kb或者10kb）寻找TF的motif富集。如果miRNA上游选定区间内出现特定TF的motif，则将这个TF作为该miRNA的预测调控子，即这个TF能够激活该miRNA的转录。这种TF-miRNA转录调控模式在基因调节中发挥作用。如图10-10所示，A图中p53同时激活mir-122a和下游靶基因*CCNG1*的转录，同时mir-122a靶向结合CCNG1的3' UTR，抑制*CCNG1*的翻译或者降解其mRNA，这个调控模式中p53和mir-122a对下游靶基因*CCNG1*显示相反的调节效应，p53转录激活mir-122a关系在控制*CCNG1*的表达中显示出抵消性地作用方式。研究发现这种不一致的TF-miRNA调控结构对于维持细胞中一些关键蛋白的稳态具有重要的作用。此外，有研究将这种不一致调控模式称为一种缓冲机制，p53和mir-122a共同作用于下游靶基因能够有效地缩短应答延迟，进而产生有效的噪音缓冲，以及精确的识别和维持细胞的稳定状态。另一种调控模式如图10-10中B图所示，p53调控mir-106a的转录和抑制*RB1*的转录，同时*RB1*是mir-106b的靶基因，说明p53和mir-106a协同作用于*RB1*的表达。其中，p53调节mir-106a转录表现为一致性地抑制*RB1*基因，这种调控模式在基因表达调控网络中起促进作用。

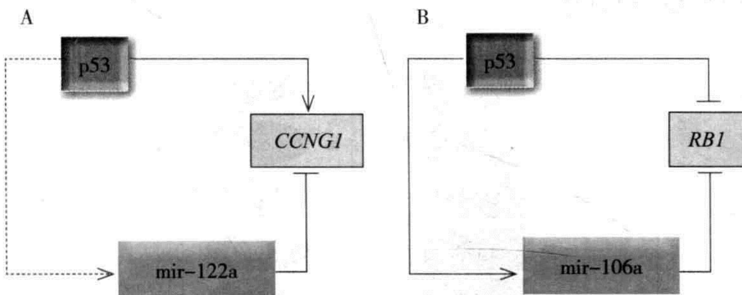


图10-10 **p53诱导的miRNA转录调控环**

大量证据表明染色质状态能够介导调控信号和DNA通道等。研究者们在分析染色质状态过程中发现，特定组蛋白标记可以准确地刻画基因组上的调控元件，如启动子、增强子、绝缘子，而这些调控元件与TF的绑定及转录的起始、延伸密切相关。例如，基因组上明显的H3K4me3信号标志着基因的转录起始，这类信号主要分布在基因的启动子区域，而H3K27me3标志着基因的转录抑制，它倾向于分布在基因组上的失活区域。因此，目前除了在miRNA前体上游一定区域内直接扫描TF的motif算法，许多研究者利用全基因的单个组蛋白标记或者整合多个特定组蛋白标记来系统地识别基因组上的调控元件，如miRNA的启动子结构。这类方法相比前面的方法，它的优势在于利用了新一代测序数据（CHIP-Seq）检测的组蛋白标记，能够更加精确地定位TF的绑定位点。此外，CHIP-Seq检测的组蛋白标记谱对于基因组注释，染色质状态检测及关联基因活性的确定提供了可靠的分析方法。这种方法能够预测特定细胞条件下处于激活状态的调控元件和它所作用的靶基因，从而能够检测到具有活性的调控关系。同时，利用CHIP-Seq检测的特定TF的全基因绑定图谱，将TF的绑

定信号映射到预测的miRNA的活性启动子上，从而检测到该细胞条件下激活的TF-miRNA调控关系。例如Mason等人利用鼠类多个细胞系（包括胚胎干细胞和成体细胞）中检测的全基因H3K4me3信号识别了miRNA的启动子。基于CHIP-Seq检测的4个核心TFs（OCT4、SOX2、NANOGHE、TCF3）的绑定图谱预测了具有显著富集TF绑定位点的miRNA启动子，发现4个TF共同作用于胚胎干细胞中高表达的mir-290-295簇的启动子，进而激活了该miRNA簇的转录。同时OCT4、SOX2、NANOG和TCF3还激活下游信号通路和转录调控通路中的一些重要蛋白。结果表明在鼠胚胎干细胞的发育分化过程中，miRNA参与调节了核心的转录调控网络。如图10-11所示，OCT4、SOX2、NANOG和TCF3直接作用于下游Lefty1和Lefty2的启动子，这两个基因都在胚胎干细胞中呈现高表达状态。Mir-290-295簇的启动子上也显著富集OCT4、SOX2、NANOG和TCF3的绑定，同时，它在转录后靶向基因Lefty1和Lefty2。因此，胚胎干细胞中的核心TF促进了Lefty1和Lefty2的表达。同时在转录后水平，一组被核心TF激活的活性miRNA通过与基因3' UTR结合进而微调这些信号通路中的蛋白质表达。这个TF-miRNA调控结构与上面的不一致性调节模式相似，它同样能够维持胚胎干细胞环境的稳态，同时TF和miRNA对下游靶基因的协同调控模式更加细致地揭示了胚胎干细胞中介导其增殖和分化的复杂作用机制。

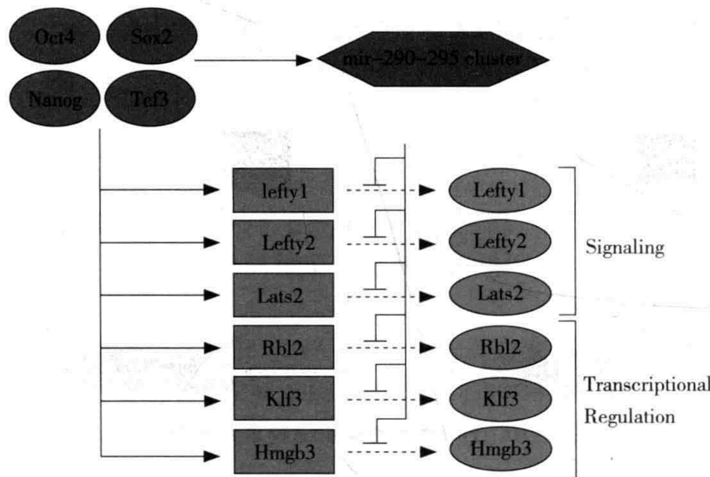


图10-11 miRNA参与胚胎干细胞的核心调控环路

二、miRNA功能协同网

miRNA与miRNA之间的协同作用，是最近几年才逐渐被人们所关注的科研方向。与不同基因之间的相互作用一样，不同的miRNA在功能上同样有着相当多的联系，人们也正在努力地尝试将其全面、准确地刻画出来。而通过网络这样一种直观、可视的方式，将相互之间存在联系的miRNA以“边”的形式连接起来，从而构建出miRNA的协同网络，显然是比较容易为人们所接受的手段之一。事实上，目前对miRNA协同作用的研究所采用的方法，往往也正是通过构建miRNA协同网络来进行数据分析和结果描述。

由于单个miRNA所对应的靶点通常很多，往往有几十甚至数百个，因此许多不同的

miRNA之间都存在着大量相同的靶点。再结合"生物体内存在大量的共表达的miRNA"这一现象，人们自然就会猜测：不同miRNA之间是否存在某种功能上的联系？进而又会想到：调控同一基因的不同miRNA之间是否存在着某种协调的机制，使得它们能够共同作用以维持基因的正常表达水平？答案是肯定的。得益于miRNA的敲除、转染方法，大量的miRNA敲除（或转染）实验都显示出：大多数的单一miRNA的异常，很难真正影响到细胞或生物的表型，即单一miRNA的异常对于表型的直接作用往往是微乎其微的。人们推测这种情况正是由于miRNA的功能冗余性造成的。单个基因往往被大量的miRNA同时调控，所以单一miRNA的异常，很难真正影响到某个特定靶向基因的表达。因此，寻找那些对同一生物学功能同时起作用的miRNA的重要性就凸现出来了。只有将这些miRNA同时考虑进去，人们才能更确切的了解某一特定miRNA的异常对某种表型的影响有多大。以某种疾病为例子，疾病状态下的差异表达miRNA很多，然而只有了解那些对疾病表型真正起作用的一个或几个miRNA，才能从miRNA调控的角度上更加明确该疾病的发病机制，并针对性的设计出更有效的治疗方案。这一切都基于人们对于miRNA之间协同作用的深入了解，同时也正是人们对miRNA的协同作用进行深入研究的意义所在。

目前，在分析miRNA的协同作用时，有两种比较常见的思路。传统的思路是在miRNA的表达谱这一层面上进行研究，分析在表达上具有一定联系的miRNA是否共同参与某一生物学过程；而后，随着人们对miRNA靶基因研究的逐渐深入，人们又可以从miRNA的靶基因的功能与联系这一角度，对miRNA间的功能联系也进行一定程度上的描述和解释。由于从表达水平的层面对miRNA的功能联系进行分析，更加类似于传统的表达谱分析，所以当Xu等人着手构建miRNA的协同网络时，就是从不同miRNA的共同靶基因是否具有功能一致性这一角度入手的，从而能够较为全面的刻画出miRNA间的协同关系，并对miRNA间的协同作用进行更为深入的阐述。在本小节我们从一个基础的例子入手，简要的分析以共同的靶基因为基础，构建miRNA协同网络的过程。

【例10-2】基于共同靶基因构建miRNA协调网络

数据：在构建miRNA协同网络时，对于最基本的，基于共同靶基因的方法来说，我们需要3种数据：miRNA集、mRNA集以及它们的靶向关系。

网络构建：在构建网络时，我们先构建一个miRNA和mRNA的连接矩阵。即分别以mRNA和miRNA为矩阵的行和列，以1和0分别描述对应的miRNA和mRNA是否存在靶向关系。当然，如果你采用了可以给出靶向关系权重的miRNA靶预测算法，我们仍然可以用阈值的形式将这种权重新划分为1和0，也可以根据需要直接使用这个权重来组成连接矩阵。如果我们称这个矩阵为矩阵A，那么A和其转置A^T相乘，就得到了一个新的，行和列均为miRNA的矩阵C，以描述miRNA与miRNA之间是否有共同的靶基因以及共同靶基因的个数（或者权重总和）是多少。这一矩阵也可以用下式表示：

$$C_{jk} = \sum_i A_{ij} A_{ik} \tag{10-4}$$

其中，C_{jk}表示矩阵C中任意一对miRNA j和k所对应的值；i代表靶向关系中所包含的全部基因；A_{jk}和A_{ik}分别表示对于任意的基因i，miRNA j和k是否与其存在靶向关系，或这种关系的强度是多少。而对于所有C_{jk}不为0的miRNA对，我们可以将其连接起来并构成一个初级的miRNA协同网络。诚然，具有相同靶基因的miRNA对都有相互协同作用的可能，但对于不同

的miRNA对来说，这种可能性的差异是很大的。接下来，我们要做的就是以某种规则，在这个大的协同网络中挑选协同关系更强烈的miRNA对所形成的子网，或者说将大网中关系较弱的miRNA对删除。

统计量和阈值的选取：显然，为了衡量miRNA对互作关系的强弱，miRNA对之间共同靶基因的数目或权重之和q是最容易想到的统计量（图10-12）。对于那些至少含有q个共同靶基因（或者共同的靶基因权重总和达到q）的miRNA对来说，我们就可以将其连接起来；而共同靶基因数小于q的miRNA对的协同关系将被删除。这样我们就可以构建出一个具有较强协同关系的miRNA协同网络。我们的下一个目标就是找出“最优”的q^*，以满足不同的需求。

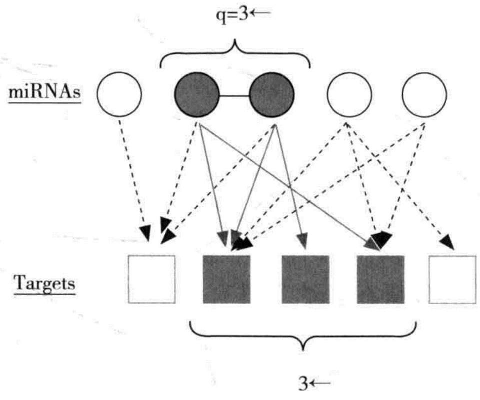


图10-12 q统计量示意图

在这个例子中，采用了保护网络中的hub节点（连通度较高的节点）的思想来对q^*进行优化。我们知道随着q值的增加，miRNA协同网络将逐渐变得稀疏和“破碎”。如果我们定义“网络碎片数”即不联通子网数为N的话，随着q的增加，N的值也将渐渐升高。然而，随着q的增加，N的变化速度并非是恒定的。在N变化最剧烈时，说明大量的具有较高联通度的节点的边都被破坏了，才导致了网络整体联通性的剧烈下降。我们可以定义此时的q值为最优值q^*，它代表了大量的与高连通度节点相连的边的q值。为了求出q^*，我们可以尝试不同的q以给出N和q的关系，进而拟合出N和q的方程。基于这个方程，按$\frac{a}{d_q}N_q$求导即可得出我们所需要的q^*值。此时，在“破碎”的miRNA协同网络中，我们就可以找到一个最大的，或几个较大的子网用于后续研究了。

当然，这是一种最直接的选取统计量和阈值的方法。大量结合其他数据的方法都已经逐渐被人们采纳了。比如我们在构建统计量的时候，完全可以加入靶基因本身的功能信息（使用基因的功能注释方法）。即，我们在考虑miRNA对的共同靶基因时，仅仅考虑那些具有相同功能的基因群。如果一对miRNA的共同靶基因的功能完全不同，即便它们的q值很大也不会被认为是具有协同关系的。事实上，无论多复杂的miRNA协同网络构建方法，其根本的区别也无非是在这一步构建不同的统计量，并选取合适的阈值而已。还有很多可行的思路能够用于统计量和阈值的可供人们选择，甚至不需要加入新的信息。比如在选取q^*的时候，是否可以考虑到随机情况下q与N的关系？事实上我们要做的也正是将那些假的、弱的协同关系剔除。那么这些关系随着q的增加而减少的速率和真实协同关系的速率显然是不同的，

那么由此产生的N的增加速率也是不同的。因此,根据随机情况下q-N曲线与真实情况下q-N曲线的差别,我们同样可以找出更优的q^*。由于方法较为复杂,这里就不再详细叙述了,仅供有兴趣的读者参考。

最后需要注意的是,在实际的工作中人们往往是将靶基因信息和miRNA表达信息联合起来考虑的。我们可以应用特定状态(比如癌症)下的表达谱建立特定条件下的miRNA协同网络。虽然这种做法不能显示出miRNA普遍存在的功能联系,但它更有针对性的描述了特定情况中miRNA的复杂调控和协同关系。在深入了解这一疾病状态,并找出关键性的miRNA群的过程中,它给人们带来了巨大的方便。此外,还有很多的思路和方法都可以应用于构建miRNA的协同网络,比如考虑靶基因的表达或功能等。对于这样一个新兴的科研方向来说,还有大量的问题需要人们去解决,还有无数的难关等待着人们去攻克。

【例10-3】利用TargetScan预测的靶基因数据构建miRNA功能协同调控网络并分析疾病miRNA的拓扑性质

数据准备: 构建miRNA-miRNA功能协同调控网络需要三种类型数据: miRNA—gene调控关系数据、基因功能注释数据、蛋白质互作数据。其中miRNA调控数据来自TargetScan5.1,下载了保守和非保守靶点数据。本文认为context score≤-0.3才是潜在的靶点,获得了185773条调控关系,涉及676个miRNA和15 829个基因。蛋白质互作数据来自HPRD(HPRD_Release_8_070609),这里我们只分析最大组分。预处理以后,最大组分包含8556个蛋白,33 762个互作。基因的生物过程功能(简称BP)注释数据来自Gene Ontology数据库,下载地址: http: //www.geneontology.org,时间2009-11。依据前人的研究成果,我们只考虑BP中那些位于第四层或者更深层次的节点。

构建miRNA-miRNA功能协同调控网络: 当我们对数据预处理后,就可以整合这三种类型的数据来识别miRNA功能协同调控对,图10-13表述了我们方法的流程。首先,对每个miRNA对,我们将它们共调控的靶基因作为一个靶点子集,然后识别这个靶点子集中候选的功能模块,这些候选功能模块的寻找是通过在GO中BP本体论的功能富集实现的。累积超几何分布被用来计算该靶基因子集在所有被考虑的BP功能类上的功能富集程度。当miRNA对至少调控一个候选功能模块时,我们用蛋白质互作网络中两个拓扑特征来过滤出功能模块,限制如下: i)每个靶基因到模块中其他靶基因的最小距离都不大于给定的阈值$D1$;ii)模块的特征路径长度要小于$D2$并且和随机情况比较要显著小。这里,我们产生了1000个随机网络。作为严格的对照,随机网络是通过保持每个蛋白的直接互作邻居不变,通过用边扰动的方法实现。总之,功能模块需要满足三个条件: 被miRNA对共调控,富集在同一个GO功能类中,在蛋白质互作网络中距离近。这里,如果一对miRNA显著共调控至少一个功能模块,我们就定义这两个miRNA是功

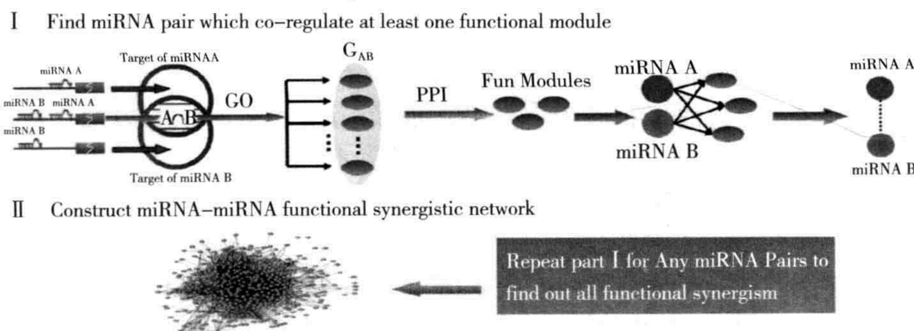


图10-13 功能协同miRNA对的识别流程

能协同作用的。最后，所有miRNA协同作用对形成miRNA–miRNA功能协同调控网络（miRNA–miRNA functional synergistic network，MFSN）。其中，节点表示miRNA，边表示它们之间的协同作用。

miRNA–miRNA功能协同调控网络的结构特征分析：理论上讲，在所考虑的miRNA对（676*675/2）和3894生物过程类之间共有888 416 100个可能的衡量功能富集的*P*值。当给定功能富集的显著性水平$P<0.05$时，我们检测到miRNA对和候选功能模块之间的472 573调控关系。经过蛋白质互作网络中两个拓扑限制和特征路径的显著性水平设为$P<0.001$，13687功能模块被473个miRNA协同调控，其中一对miRNA可能调控多个不同的功能模块。这些miRNA间有2937个非冗余的协同模式。从图10–14我们发现几乎所有的miRNA都连接在一起，并且有一个小的半径（2.8691）。我们用复制模型产生随机网络，这些网络的平均半径和真实的类似（2.8722 ± 0.1332），也是小世界网络。但是MFSN网络还展现出紧密的邻居关系，平均聚类系数达到0.2747，比随机网络的显著高（0.0684 ± 0.0151）。这是因为miRNA的直接邻居，即功能协同对象，也倾向有协同作用。小世界网络的这种紧密的邻居特征是特别有意义的，因为它能用来预测新的协同作用，就像以前预测蛋白质互作中那样。另外，在MFSN中，只有一些miRNA有相对多的协同作用，大部分miRNA只有很少的协同邻居。检验这个MFSN网络的度分布发现，该分布服从斜率为0.7902（拟合优度R2=~0.9264）的幂分布，表明MFSN是一个无尺度（scale-free）网络。

我们分析了MFSN网络的模块和社区特性。这里，我们定义一个miRNA功能协同模块为一个最大完全子图（clique）。MFSN中所有的模块（或社区）都有独一无二的miRNA组成，但是也允许同一个miRNA或者同一条边出现在多个模块中。我们统计了每种k值对应的模块数目和模块中miRNA占所有miRNA的累积分布。结果发现，随着k值的增加，模块的数目急剧下降。总共有77.51%的miRNA至少包含在一个模块中。

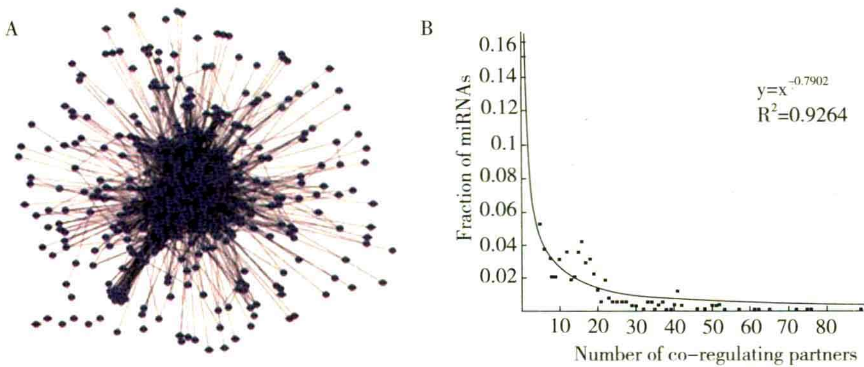


图10–14 MFSN网络图和度分布

A. 表示MFSN，其中每个节点代表一个miRNA，边代表协同作用；B. 表示度分布

我们解释这种现象可能是下面原因导致的，即miRNA完成某种特定调控时，是以小的团的方式完成的，而不是单个发挥作用或者以大模块的方式。因为同一个家族的miRNA倾向有相似的功能或者参与同种疾病，我们进一步调查是否同一家族的miRNA倾向出现在同一模块或者社区中。一共有70个miRNA家族至少包含两个miRNA，其中60%的家族完全被包含在至少一个模块中，在社区中这个比例高达65.71%。因此，同一家族的miRNA确实倾向功能协同。

总之，miRNA功能协同调控作用相互交织，形成一个复杂的miRNA功能协同调控网络。该网络具有小世界和无尺度的特性。因此，类似于许多大的网络，MFSN的无尺度特征暗示着该网络

并不是一个随机网络，而是被一个核心的组织原则集合刻画，这些原则能够使其区别于随机连接的网络。已有研究证明一个干扰在小世界网络中散播所花费的时间接近于理论上任何有相同点和边的网络中所用时间的最小值。因此，小世界网络允许miRNA协同作用能快速响应干扰。

疾病miRNA的拓扑特征：目前越来越多的研究给出了特定疾病相关的miRNA，有些科研院所构建了专门的数据库收集这些关联信息，例如哈尔滨工业大学开发的miR2disease和清华大学开发的HMDD数据库。我们基于miR2disease数据库提供的miRNA疾病信息做了初步探讨。基于疾病中miRNA差异表达的检测方法，我们获得两类不同可信度的疾病数据。一类是我们从数据库中所能获得的全部疾病miRNA信息；另外一类是前面数据的子集合，由那些低通量实验证实的疾病miRNA组成，比如“Northern blot”和“qRT-PCR”。结果发现第一类包括236个miRNA，它们参与了108种疾病，第二类包括164个miRNA，涉及94种疾病。

我们分析疾病miRNA在miRNA-miRNA协同调控网络中的拓扑性质，如图10-15。我

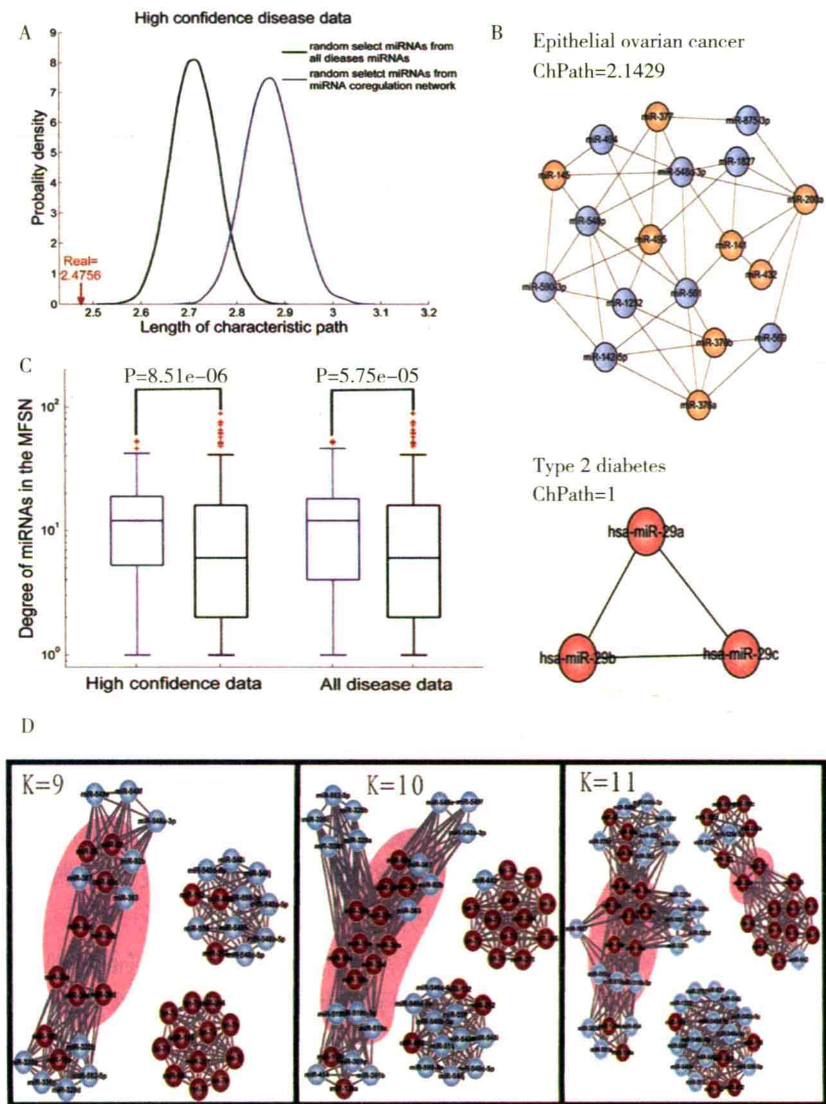


图10-15 疾病miRNA的拓扑特征

们发现同一疾病的miRNA在协同调控网络中距离和随机比较显著的近，暗示着同一疾病的miRNA调控相同或者相似的功能，存在聚集现象。比如，卵巢癌的上皮细胞中用低通量实验已经识别14疾病相关miRNA，我们发现其中有8个miRNA至少由一条功能协同调控，它们的特征路径为2.1429。另外，结果发现miRNA协同调控的数目和其调控功能模块的类型是显著正相关的，且疾病miRNA有更多的协同作用。因此，疾病miRNA有更高的功能复杂性。疾病miRNA还倾向定位在大的miRNA-miRNA协同调控模块中，特别是这些模块的交叠处，表明疾病miRNA倾向是协同调控网络的全局中心，对不同或相似生物过程起到衔接作用。归纳出疾病miRNA在协同作用网络中的拓扑特征可以开拓我们对miRNA致病机制的理解，还可以用来预测新的疾病miRNA。

总之，miRNA-miRNA功能协同调控网络不仅可以用来深入理解miRNA的转录后调控模式，还为探索疾病miRNA的性质提供了一个新的视角。

三、miRNA调控不同分子网络

生命是存储并加工信息的复杂过程，以往孤立地研究单个基因及其表达的变化往往不能够确切地反映生命现象的本质规律。这也是分子生物学家在从基因组序列到蛋白质结构层面获得了海量的数据之后，面对信息整合感到迷茫的真正原因。2003年，Bray提出了关于生物分子网络整合的建议。聪明的科学家从社会科学那里借鉴了用来处理数据的“网络”来系统地探究生命现象。简而言之，活体细胞就是一个网络，细胞中的一切分子都存在着普遍的生物联系，细胞与细胞之间也存在着信息的传递，因而构成了复杂的生物分子网络系统。依据中心法则，我们知道RNA只是负责将遗传信息从基因组传递给蛋白质。而最近发现的非编码miRNA却挑战了这一法则，它被认为是信息流的调控者之一，并且有越来越多的证据显示miRNA的调节作用涉及生命过程的方方面面。

我们已经知道通过靶预测算法得到的miRNA的靶基因种类相当广泛，包括信号蛋白、代谢酶、骨架蛋白和转录因子等。miRNA靶基因的多样性和丰富性暗示着miRNA在和它的靶基因形成复杂网络的同时，必然和其他的分子网络（如转录调控网络）存在复杂的交互作用。因此，miRNA通过调控分子网络行使功能的观点是合理的，我们有必要从系统水平上去解析miRNA是如何参与细胞调控过程的。

基因调控网络、信号网络、蛋白互作网络和代谢网络是目前研究比较广泛的四种分子网络。其中，基因调控网络描述转录因子和编码蛋白基因间的调控关系，其包含细胞中全部生物过程的基因调控信息。而蛋白互作网络则包含了蛋白信息及其物理互作的信息。蛋白互作囊括了从基本的细胞机制（如DNA合成蛋白复合物）到细胞信号涉及的蛋白复合物等信息。简而言之，基因组范围的蛋白互作网络包含了细胞生命过程涉及的全部蛋白互作信息。因此，我们将基因调控网络和蛋白互作网络划分为“一般网络”。

miRNA通过靶向基因的3’UTR对靶基因进行功能调控。因此，将miRNA的靶基因映射到蛋白互作网络（中间需要先将靶基因同蛋白对应）或是转录调控网络，就可以将miRNA及其靶基因形成的网络同二者联系起来。对于蛋白互作网络，截至目前，酵母、大肠埃希菌及其他细菌、线虫、果蝇和人类均已进行了大规模的蛋白互作检测，使得蛋白互作数据相对完善，所以对网络中miRNA的转录后调控机制的探讨也最为细致。研究显示即使不限定蛋白

互作网络及靶数据来源，也能够得出与以往类似的结果，即转录后调控复杂度、互作复杂度及功能复杂度是正相关的。网络中的HUB蛋白更倾向被miRNA调控，互作的蛋白倾向具有类似的转录后调控模式。单个miRNA的靶基因模块化不是很明显，但加入靶基因的互作邻居后模块化显著。而对于miRNA簇，同样可以找到其显著调控的功能模块。

对于转录调控网络，其和蛋白互作网络明显不同的是，全基因组范围内的基因调控数据还不完善，依靠低通量实验获得数据比较少。因此只能在有限的数据中分析miRNA的调控规律，然后再分析预测得到的转录因子在启动子结合位点的数据，寻找miRNA的调控规律。结果显示两者有较高的一致性。即基因拥有的转录因子结合位点越多，这个基因的转录后调控机制越复杂，其越倾向被较多的miRNA靶向。反过来，基因上miRNA的靶点数目越多，其倾向于拥有越多的转录因子结合位点。也就是说miRNA的调控数目和转录因子结合位点的数目是显著正相关的。换句话说，人类基因组中，转录后水平上miRNA的调控复杂度和转录水平上转录因子对基因的调控复杂度是正相关的，即转录调控越复杂的基因，其转录越需要频繁开启，越有可能具有时空特异性表达，因此其转录越需要频繁关闭，即越复杂的转录后调控。另外，当我们对同时具有转录调控及转录后调控复杂性的基因进行功能富集（数据来自GO数据库）时，结果显著富集在和发育相关的生物过程。由于人类和啮齿类生物中关于基因的正向或负向转录调控关系缺乏，我们现在还不能分析转录调控网络的局部模块化。

代谢网络和信号网络，通常被称为“特异细胞网络”，其描述某些特定的细胞活动中信息流的行走方式。细胞代谢网络包含所有代谢反应和代谢流，而信号网络则包含信号流和信号传导过程涉及的生化反应。通常这两种信息间的交互用线性通路表示，比如代谢通路和信号通路。代谢通路是紧密交织的，因此代谢流可以通过多条通路进行传递。且有些代谢产物是被多条通路共享的，因此可以通过一条或者多条通路得到某种代谢产物。信号网络涉及细胞内和细胞间的交流以及信号蛋白对信息的处理方式。其作为高级交流系统能够完成诸如生长，细胞存活和发育等功能。

2006年，Cui等把miRNA靶数据进行信号网络映射，揭示了miRNA调控人类信号网络的一般规律。通过信号蛋白类别的划分，依据各类别中被miRNA靶向的信号蛋白的比例，分析miRNA靶向的倾向性。结果显示miRNA倾向于靶向信号流下游的核蛋白成分（大部分为转录因子）。而通过功能将信号蛋白进行分类，继而研究miRNA靶向的倾向性时，Cui等主要探讨了连接蛋白，无酶活性，通过和上游及下游信号蛋白紧密互作来实现信号传递。结果显示，miRNA倾向于靶向高连接组（下游多于4个信号蛋白），并通过调控连接蛋白下游成分的浓度而精确响应各种刺激，这和miRNA的高时空表达特异性也是一致的。

网络模体，作为信号网络中一种普遍存在具有简单结构的网络单元，不同的网络模体代表不同的信号传递模式。利用Mfinder程序提取网络中三到四个节点的模体，依据模体中被miRNA靶向的蛋白的比例研究miRNA靶向的倾向性。结果显示，miRNA倾向靶向正向调控的模体。正向调控的模体中，任何成分的噪声或者波动都容易被放大，从而使生物系统的状态发生转换。而miRNA的负向调控能够增强其对这种放大作用的过滤或者缓冲，从而精确调定和维持细胞稳态。网络模体通过共享的信号蛋白相互连接组成的更大网络结构，称为网络主题（theme）。miRNA的功能预测可以通过寻找与其紧密关联的网络主题实现。而大部分网络主题和五个细胞机器（包括转录机器、翻译机器、分泌机器、活力机器和电荷机器）中的一个或者更多相关。但这些机器的共享蛋白，却不倾向于被miRNA调控。由于其功能

较为基础，表达水平相对稳定，不需要过多的调节，因此避免miRNA的调控。

对于代谢网络，许多代谢产物可以被不同的代谢通路共享，使得各通路间形成复杂的交互。其调控的复杂度涉及转录、转录后和翻译水平三个层面。Tibiche等系统分析了人类代谢网络中miRNA的调控模式。用KEGG数据库中获得的人类代谢通路数据以反应为中心的模式来刻画代谢网络，即有向图的模式。然后将TargetScan预测得到的miRNA的靶数据进行代谢网络映射。如果反应只包含一个酶且被miRNA靶向，则认为该反应被miRNA调控；若包含多个酶则需要跟随机（随机扰动miRNA和酶之间的靶向关系）进行比较得出miRNA对该反应的调控是否具有显著性来确定该反应是否被miRNA调控。同信号网络不同的是，网络节点的分类测度。对于代谢网络，我们将节点分为上游节点、切割点（cut point，CP）、hub节点（定义为网络中前5%出度和入度和大的节点）、中间节点（intermediate nodes，ITN）和下游节点（图10-16）。通过计算各类型节点中被miRNA靶向的百分比（对于包含多个酶的反应还需要计算其百分比同随机比较是否具有显著性）来确定其是否被miRNA靶向。结果显示，miRNA显著地不调控ITN；却显著富集在hub和CP两类型节点上。同信号网络类似，代谢网络也存在模块化的结构——代谢流，其对应某种特定物质的代谢过程。代谢流可以分为两种结构方式：线性模式和支化模式，后者又可以继续分为两种子模式：分叉型和汇集型（图10-17）。这三种模式的总数可以通过枚举网络中连接的两个或者三个反应的组合得到。然后，根据该模型是否被miRNA靶向而对其进行分类，然后分别计算三种模型跟随机比较是显著的出现还是显著的缺失。结果表明，线性模型代谢流都被miRNA靶向，同随机比其是显著出现的。这表示代谢网络中某些局部反应区域被miRNA显著调控。另外，无论是汇集型还是分叉型，不包含miRNA靶点的模式是在网络中显著出现的，至少包含2个靶点的模式在网络中是显著缺失的。对于miRNA调控倾向性的研究，则显示，miRNA广泛调控基本的代谢通

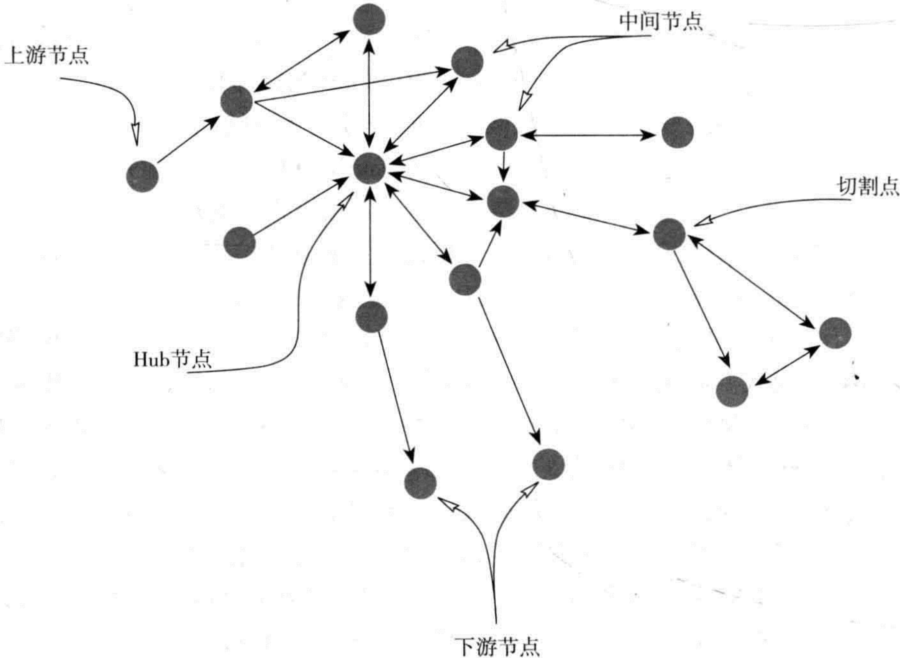


图10-16 代谢网络中五种节点类型

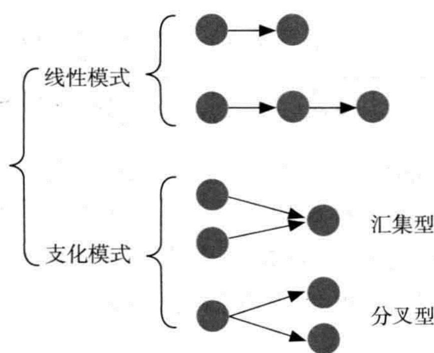


图10-17 代谢网络中代谢流的分类

路，如氨基酸合成/降解和某些脂类代谢通路，暗示着miRNA在细胞代谢中作用的重要性。

把没有入度的点称为上游节点，即该反应的底物不是其他任何代谢反应的产物；同样的，把没有出度的点称为下游节点，即反应的产物不是其他任何反应的底物。删除后能增加网络组分数的点称为切割点，而拥有较高出度和入度加和的点称为hub节点。网络中剩余的所有节点称为中间节点。

线性模式即一个反应的产物是且只是另一个反应的底物，文中仅考虑两个或者三个反应的线性模式；对于支化模式，进一步分为汇聚型，即两个反应的产物是另外一个反应的底物；分叉型则表示一个反应的产物是另外两个反应的底物。

【例10-4】基于miRNA-靶点失调网络的拓扑特征优化前列腺癌风险miRNA

这里，我们提出一种猜想，我们认为除了表达异常的miRNA和疾病相关，另外一类miRNA也应该值得注意，即对靶基因调控发生异常的miRNA。miRNA是通过对靶基因进行调控而完成生物学功能的，如果miRNA对靶基因进行异常调控，很有可能会导致癌症的发生。因此，这里我们借助miRNA在miRNA-靶基因失调网络中的拓扑特点来优化前列腺癌风险miRNA。

1. 数据准备 构建miRNA-mRNA失调网络需要三种类型数据：miRNA—gene调控关系数据、前列腺癌中miRNA和mRNA表达谱数据。其中，miRNA—gene调控关系数据是TargetScan（5.1版本）、miRanda（miRBase数据库中下载）、PicTar和DIANA-microT预测结果的并集。表达谱数据是Ambs等人检测的，从GEO数据库下载获得，其中miRNA表达谱对应的GEO编号为GSE8126，mRNA表达谱数据对应的GEO编号为GSE6956。这里，我们只关注同时检测mRNA和miRNA表达的样本，共75个，其中60个是原发性的前列腺癌样本，15个是作为对照的前列腺癌旁组织，此处认为是正常的样本。我们对mRNA的表达谱利用robust multi-array average（RMA）算法进行标准化处理，而对miRNA则直接从网上下载已经进行过标化处理的数据。

2. 构建miRNA-mRNA失调网络 我们逐次判断每对有表达的miRNA和它的靶基因之间的调控关系在前列腺癌样本和正常比较是否发生显著失调（图10-18）。首先，miRNA和mRNA的表达谱被分成两部分，分别是肿瘤样本的表达谱和正常样本的表达谱。其次，对每对miRNA i 和靶点 j，分别计算它们在正常样本和肿瘤样本中的皮尔森相关系数，并观察两者之间的差异，公式如下：

$$Dys=\left(\frac{\sum\left(M_C-\overline{M_C}\right)\left(T_C-\overline{T_C}\right)}{\left(n_C-1\right)S_{M_C}S_{T_C}}\right)-\left(\frac{\sum\left(M_A-\overline{M_A}\right)\left(T_A-\overline{T_A}\right)}{\left(n_A-1\right)S_{M_A}S_{T_A}}\right) \quad (10\text{-}5)$$

其中M和T分别表示的是miRNA i和靶点j的表达值。C是癌症样本，而A是正常的前列腺样本。n_C和n_A分别表示癌症和正常样本的个数。$\overline{M_C}$，$\overline{T_C}$和$\overline{M_A}$，$\overline{T_A}$是miRNA i和靶点j在肿瘤和正常样本中的平均表达值，而$S_{M_C}S_{T_C}$和$S_{M_A}S_{T_A}$分别表示两种类型样本中miRNA和靶基因表达值的标准差的乘积。这个Dys测度可以用来估计miRNA和靶基因在两种类型样本（肿瘤和正常）中相关性的失调程度。

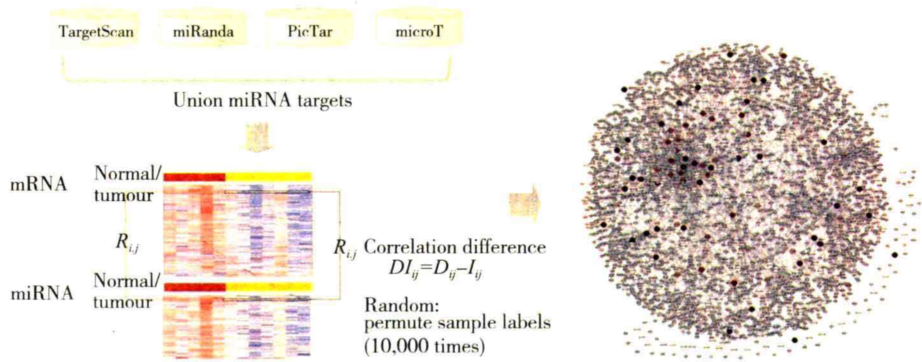


图10-18 概览miRNA–靶点失调关系的识别过程

为了确定是否两类样本中miRNA和靶基因之间的失调是显著的，我们随机扰乱所有的样本标签10 000次并重新计算值。这样，我们对每对miRNA和靶基因的关系可以计算一个经验概率。然后，我们用对所有的调控关系的概率值进行多重校正检验，当FDR≤0.01时这对关系才被认为是显著失调的。

我们识别了前列腺癌背景下的3758条显著的miRNA–靶基因失调关系，涉及274个miRNA和2511个基因。这些失调关系并不是孤立存在的，而是彼此交错形成了一个复杂的miRNA–mRNA失调网络，这里称之为miRNA–靶基因失调网络（miRNA target–dysregulated network，简称MTDN）。所以，我们可以融合上面已知疾病miRNA的拓扑性质来优化候选的癌症标记。

3. 优化前列腺癌风险miRNA　基于疾病miRNA在协同网络中拓扑性质的分析，我们发现了疾病miRNA是以模块的方式行使功能的，并且疾病miRNA倾向于位于模块的交叠处，能够更加便利信息的交流。我们发现构建的前列腺癌背景下miRNA–mRNA失调网络也是无尺度的。我们搜索miR2disease等疾病miRNA数据库和阅读文献获得了37个已证实的前列腺癌miRNA，将其作为真阳性miRNA集合；另外，我们将在正常前列腺组织中表达最低的50个miRNA作为真阴性miRNA集合。结果发现这两类miRNA的拓扑性质的确存在显著的差异：疾病miRNA有较多的失调靶基因，其中很多靶基因都是和其他前列腺癌miRNA共同失调的；还发现它们有较多的协同调控miRNA，其中前列腺癌miRNA占的比例也显著高于其他miRNA。这暗示着在前列腺癌中miRNA有聚集现象。进一步地，我们利用这些拓扑性质及表达的改变为特征构建了SVM分类器，用五倍交叉证实进行评价，结果表明该分类器准确性达到了0.8872，要明显优于单单利用miRNA表达来预测候选miRNA（图10–19）。然后我们利

用训练好的分类器来预测新的前列腺癌miRNA，发现很多已知的前列腺癌miRNA都排在前面（图10-19）。通过对得分比较高的未知前列腺癌miRNA进行功能注释，发现这些miRNA的功能和前列腺癌的发生有密切关系。例如，hsa-miR-203失调的靶基因富集了Hedgehog信号通路、细胞分化和细胞增殖等。

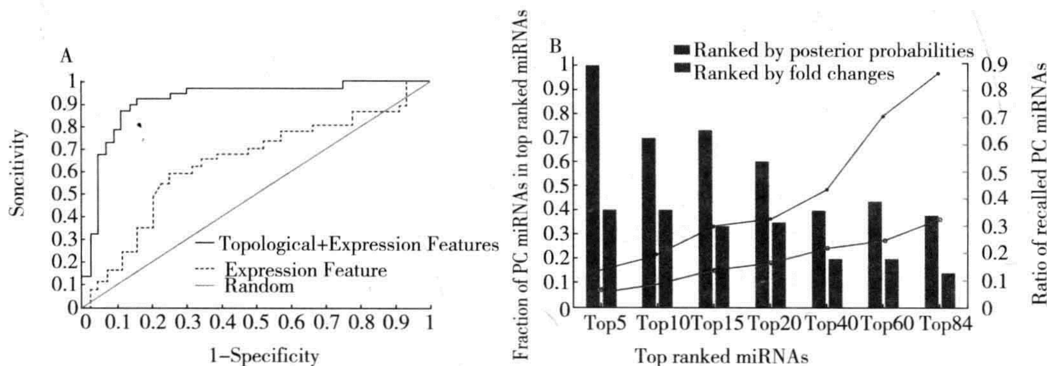


图10-19 基于拓扑特性构建的疾病miRNA分类器评价

四、miRNA调控的网络模体

miRNA是内源性的非编码RNA，在转录后水平上对其靶基因的表达起负向调节作用，进而参与多种生物学过程，如发育、分化、生长、代谢、凋亡、信号通路甚至癌症的发生和发展。很多研究发现转录因子和miRNA高度共表达，表明转录因子介导的转录调控和miRNA介导的转录后调控经常是高度协同的。为了更好地探究转录因子和miRNA的协同效应，刻画基因组规模的全局调控网络是非常重要的。以网络为媒介可以有效地分析转录因子和miRNA是以怎样的联合调控方式调节基因表达，从系统水平上揭示基因表达的调控机制，而不是仅仅从单个基因的水平。通过对调控网络的分析，很多研究发现在网络中存在着一些显著富集的调控结构，进而网络模体的概念应运而生。网络模体（network motif）被定义为网络中较小的调控回路或者结构模式，它在真实网络中出现的次数统计学上显著高于在随机网络中出现的次数。转录调控网络的网络模体首先在细菌和酵母中被发现，提供了调控网络中更局部的信息。目前，一些miRNA调控的网络模体已经被实验数据证实，例如，在果蝇的眼睛细胞中，转录因子Yan抑制miR-7的转录，反过来miR-7也可以在转录后水平抑制转录因子Yan的翻译，从而Yan与miR-7形成了表达水平的互相排斥状态，在EGFR信号触发的Yan降解状态下，使得miR-7与Yan的表达模式发生了一个稳定的差异，进而启动了果蝇眼细胞感光器的分化。在人类粒细胞分化以及线虫外阴细胞中也发现了类似的环路。即在不同物种的基因调控网络里面找到了相同的模体类型，说明网络模体是网络趋同进化的结果，并且在细胞内执行重要的生物学功能。

对于miRNA参与的复杂网络而言，包括miRNA在内的三个节点或四个节点所对应的网络模体的研究较为广泛。比如，Yu等人在人类的调控网络中，探究所有可能的至少包括一个miRNA和一个转录因子的三个节点的子图，发现了17个miRNA调控的网络模体。寻找网络模体的方法是将网络中的某一种感兴趣的调控模式与随机网络相比较，如果这个调控模式显著的富集到真实的网络，说明这种调控模式在真实的网络中起着重要的作用，即为网络

模体。具体说来，针对一个复杂有向网络而言，对整个网络进行遍历，寻找含有n个节点的子图（目前的研究主要集中在$n=3,4$），并记录每个子图在网络中出现的次数，然后将实际的网络随机化，采用在实际网络中寻找子图的方法遍历该随机网络，并记录遍历随机网络得到的n个节点子图的出现频数，如果实际网络中出现的次数显著比它在随机网络中出现的次数大，那么这样的n节点的子图模式就是一种网络模体。那么对于随机化的方法，为了提高随机网络与实际网络的可比性，往往在随机化实际网络的时候保持实际网络中节点的出度和入度不变。

目前，很多研究中都发现了几种广泛存在的网络模体。

1. 前馈环（Feed-forward Loop） 即上游转录因子调控一个靶基因表达的同时，还作用于下游的一个miRNA，然后与这个miRNA协同调控下游靶基因的表达。也就是上游的转录因子调控下游的靶基因是通过两个途径同时控制，一个途径是转录因子直接作用于靶基因，另外一个途径是通过调节一个miRNA，间接的作用于靶基因。按照直接途径和间接途径对靶基因效应的一致性与否，把前馈环被分为两种，即Ⅰ型miRNA前馈环（Incoherent feed-forward），与Ⅱ型前馈环（coherent feed-forward）（图10-20）。

在Ⅰ型miRNA前馈环中，上游转录因子激活（或抑制）靶基因的同时激活（或抑制）一个miRNA的转录，在转录后水平上抑制这个靶基因的翻译过程。研究发现在miRNA表达的细胞类型中，Ⅰ型miRNA前馈环可以启动这个miRNA靶向的基因，使其处于高表达水平。这表明miRNA在Ⅰ型前馈环中对其靶向的蛋白起到微调控的作用，维持蛋白的表达在一个正常的功能范围。因为真核细胞中基因的转录经常是处在一个噪音环境中的，所以相应的mRNA的拷贝会受到一定的波动，除此之外，其他的一些因素比如mRNA的降解和蛋白质的翻译也会随机的波动。更重要的是，这些波动会适当的沿着调控网络蔓延开来，如上游转录因子的波动会造成下游基因的表达产生波动。那么在Ⅰ型前馈环中，任何导致上游转录因子偏离其稳态的干扰信号，会沿着调控网络以相同的干扰趋势引起下游靶基因和miRNA的表达偏离，此时由于Ⅰ型前馈环中上游的转录因子调控靶基因的两个途径的效应是相反的，所以会使得有miRNA介导的间接途径在另一个相反的方向去缓冲外界干扰，维持蛋白表达免受波动。例如，图10-20所示在果蝇细胞中转录因子Atonal 激活E（spl）同时激活miR-7抑制E（spl）的翻译，缓冲外界干扰，维持果蝇眼睛正常发育，进而决定了感觉器官的命运。

Ⅱ型miRNA前馈环（coherent feed-forward）是转录因子通过直接途径和间接途径控制其靶基因的表达，且这两个途径方向是一致的，即靶基因的表达同时被激活或抑制。这种网络模体可以使细胞有效激活（或抑制）那些在转录水平上漏掉的靶基因。例如，c-Myc/E2F/miR-17-92环路（图10-20中Example），转录因子c-Myc启动细胞周期进程，需要激活E2F转录因子家族以及mir-17-92簇，同时被激活的E2F家族还可以作用于mir-17-92簇的启动子区域进一步激活其表达。而且这个Ⅱ型miRNA前馈环被嵌入到一个更加复杂的环路当中，因为mir-17-92簇成熟后还可以反过来抑制E2F家族，这个环路是一个Ⅰ型前馈环和一个Ⅱ型前馈环嵌套的一个结果，揭示了c-Myc同时激活E2F家族的转录和抑制其翻译，从而精密地控制了分化信号。

2. miRNA参与的反馈环（feedback loop） 如图10-20所示，被分为coherent反馈环和incoherent反馈环。Coherent反馈环是miRNA与转录因子之间对彼此的调节效应是相同的（即互相抑制或者互相激活），导致了转录因子与miRNA相互排斥的表达（互相抑制），或者是表

达水平的双稳态（互相激活），当一个瞬时的信号改变了这种模体的表达状态，当信号消失后模体的状态不会恢复，即产生了不可逆的状态改变。如图10-20所示，在人类的造血干细胞中，mir-233和NFI-A形成了一个coherent反馈环控制粒细胞性分化。在未分化的细胞中，mir-233处于低表达，NFI-A处于高表达状态，然而在视黄酸信号作用下，mir-233表达水平上调并借助mir-233和NFI-A形成的coherent反馈环导致NFI-A受抑制，进而削弱了NFI-A对mir-233的抑制作用，最终促进了骨髓系细胞的分化。Incoherent反馈环，转录因子和miRNA以相反的方式互相调控，功能是微调基因的表达并且维持转录因子和miRNA表达水平的相对稳定。在Incoherent反馈环依赖于一个输入信号，使得miRNA与转录因子的表达呈现振荡状态。例如，在线虫外阴细胞中，Notch信号蛋白可以激活miR-61的转录，而miR-61反过来抑制Notch信号蛋白的翻译过程，从而稳定了外阴细胞的状态（图10-20）。

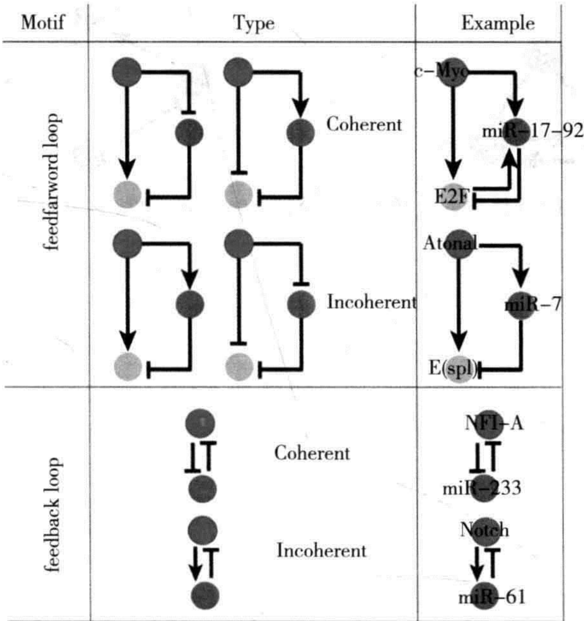


图10-20 miRNA参与的coherent前馈环、incoherent前馈环与反馈环示意图及其实例

红色圆圈代表转录因子，绿色的圆圈代表miRNA，蓝色的圆圈代表靶基因

3. 被miRNA调控的反馈环（regulated feedback loop） 如图10-21所示，这种网络模体由两个转录因子和一个miRNA组成，其中两个转录因子互相调控，并且共同受到一个miRNA的靶向。这种环路可以将一个短暂的刺激信号传递到调控环路之中，引起一个稳定的不可逆的反应状态，这种状态对于与发育相关的调控关系尤其重要。另一方面被miRNA调控的反馈环与普通的反馈环相比较，可以阻止偶然性的激活事件，因此为发育过程提供一个稳定的环境。

4. 混合调控环和间接前馈环 混合调控环是由上述提到的模体相互嵌套得到的复杂网络模体，如图10-20例子所示，c-Myc/E2F/miR-17-92环路，是由一个Ⅰ型前馈环和一个Ⅱ型miRNA前馈环嵌套得到的一个更加复杂的环路。间接前馈环由上游的转录因子A激活下游的一个转录因子B去激活一个miRNA，而这个miRNA同时跟转录因子B去调控靶基因。

5. 双重的负反馈环（double-negative feedback） 即一个miRNA通过另一个miRNA参与

的途径间接调控自身的表达水平，如线虫的味觉感受器分为两种ASEL和ASER，分别分泌不同的化学受体来感受和应对外界的输入，但是这两种感受器是由同一种前体细胞（ASE神经元）分化而来，lsy-6和die-1的高表达决定ASEL的稳定状态；类似的*mir-273*和*cog-1*的高表达决定ASER的稳定状态。研究发现ASE神经元细胞分化命运的决定是凭借两个miRNA（lsy-6、mir-273）和它们靶向的转录因子（die-1、cog-1）所形成的双重负反馈环来完成的。如图10-22所示，*mir-273*可以调节*lsy-6*的表达是借助于其直接靶die-1，同时*lsy-6*也可以通过调节其直接靶cog-1，去控制*mir-273*的表达，当一个外界的信号输入到这个双重负反馈环中，借助于双向的调节使得这四个因子的表达状态稳定下来，从而不可逆地决定ASE神经元细胞的分化命运。总言之，miRNA被封装到双重的负反馈环，为miRNA如何最终决定细胞命运提供了一个机制。

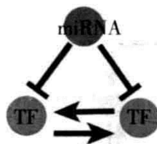


图10-21 被miRNA调控的反馈环

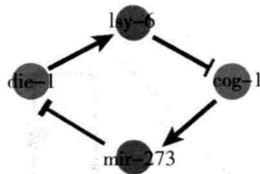


图10-22 mir-273参与的双重的负反馈环

第四节

miRNA多态和复杂疾病

Section 4 miRNA Polymorphism and Complex Disease

目前，HapMap计划在人类基因组中识别了大约一亿个SNP位点，这意味着平均每100到300个碱基对就存在一个SNP。虽然大部分SNP处于一种沉默状态，但是流行病学研究显示基因序列变异与癌症发病风险存在一定的关联。这些多态通常以插入、缺失、扩增或者染色体异位的形式出现，当它们位于miRNA基因内部、加工处理过程中或靶位点上面时，则可能影响癌症的发病风险、治疗效果以及预后。SNP作为人类基因组中一类新的功能多态，不仅能够影响miRNA的生成和表达，还能够影响miRNA与靶基因的结合从而影响靶基因的表达，最终导致miRNA的功能获得或缺失。例如，若miRNA多态使得癌基因的表达上调，便可能导致肿瘤的发生。同时，目前的一些研究也显示，miRNA多态位点与疾病的演进以及宿主对药物的反应之间存在紧密的关联。miRNA多态主要划分为四类：位于miRNA基因内的多态、位于miRNA靶点的多态、miRNA合成机制中的单核苷酸多态和表观遗传调控的多态。

一、miRNA基因内部多态

成熟miRNA通过与mRNA的3' UTR区域结合从而实现对众多靶基因进行转录后调控。新近的研究显示，miRNA种子序列内的单个碱基突变就能够消除miRNA对其靶基因的抑制作用。同时这种miRNA序列内的多态对miRNA 自身的转录、形成、输出和调控也具有重要的作用。Matthew A.Saunders 等人对人类474个miRNA进行系统分析时发现，其序列内的SNP密度低于侧翼序列内的SNP密度。研究发现65个SNP落入49个pre-miRNA内，平均密度约为1.3个SNP/kb。研究人员通过对pre-miRNA不同的功能域进行划分，进一步发现其中有3个SNP落入了种子序列。Duan等人在研究miRNA基因内SNP的分布时，也得到了相似的结果。这意味着miRNA基因内存在SNP多态。

miRNA生物合成的不同阶段需要不同的蛋白和蛋白复合物的参与，包括RNA聚合酶Ⅱ、Drosha/Pasha、Exportin-5/Ran-GTP、核孔复合体、Dicer和RISC复合物等。如果存在于pre-miRNA上面的多态影响了其与这些蛋白的结合，那么这些多态将会影响相应miRNA的生成或使得其表达下调。而位于miRNA基因内的多态则可以通过自身或是参与miRNA形成的蛋白来影响miRNA的合成与成熟，从而导致新的miRNA的生成和靶基因的识别，最终使得miRNA的功能缺失或是获得进一步影响疾病易感性和药物敏感性的功能。基于目前miRNA多态的研究成果，我们将miRNA基因内多态分为如下三类：

(一) 位于pri−miRNA和pre−miRNA基因序列内部

Saunders, M.A.和Duan等人利用生物信息学的方法进行研究,发现在pri/pre−miRNA内部存在单核苷酸多态。这些多态会影响成熟miRNA的表达以及miRNA与其靶基因的结合,甚至还会影响疾病的发病风险,同时也可能产生新的miRNA。Wu M.和Calin G.A.等人发现位于let−7e和mir−16前体内的多态能够降低其成熟miRNA的表达。而Duan等人研究发现*miR-146a*的前体(pre−miRNA)内存在的一个常见SNP(rs2910164),其C等位可以增加*miR-146a*的表达。Hoffman A.E., Tian T.和Peng S.等人发现,位于*mir-196a-2*前体(pre−miRNA)内的SNP(Rs11614913)能够增加罹患乳腺癌、肺癌和胃癌的风险,特别是纯合的CC基因型多态。同时该多态还能够影响*mir-196a-2*的成熟及其成熟miRNA与靶基因的结合。先前的研究显示,前体以及成熟miRNA序列内的变异能够影响miRNA的生物合成,在CC基因型的样本中*miR-196a-2*的表达水平要低于TT表型的样本。在肺癌中,该多态也预示着一个较差的预后,这也是第一次提出miRNA相关联的SNP与癌症的演进相关。*miR-146a*前体会产生miR−146a(正链)和miR−146a*(负链)两种miRNA。而位于miRNA−146a前体的SNP多态(rs2910164)不仅会影响*miR-146a*的表达,同时也会导致mIR−146a*C和miR−146a*G两种miRNA的生成。该研究显示, pri/pre−miRNA内的多态能够导致新的miRNA的产生。通过新miRNA对靶基因的影响,该多态引起的改变可能与多种疾病的发病风险相关。miR−146a*C能够调控基因*PTC1*,而miR−146a*G调控基因*IRAK1*。位于该多态位点的C等位能够影响*miR-146a*调控的乳腺癌相关靶基因*BRCA1*和*BRCA2*的表达。换句话说,该多态位点CG基因型罹患乳腺癌的风险要低于CC和GG两种纯合基因型。

(二) 位于成熟的miRNA序列内

成熟的miRNA通过与mRNA的3' UTR区域结合从而对mRNA进行转录后调控。其与mRNA结合的区域包括两部分: miRNA的5' 端第2−8个碱基,称为种子区域,该区域要求与mRNA完全匹配; 种子区域附近的3' 端方向,允许一定程度的错配,称为3' 容错区域(3' MTR)。位于这两部分的miRNA多态能够消除、弱化或增强其对靶基因的影响,还能产生新的结合靶点。miR−146a内部的多态,消除了其介导促凋亡转录因子的功能。

根据miRNA与靶基因结合区域,可以将成熟miRNA内的多态分为如下两类:

1. 位于miRNA的5' 种子区域 Saunders等人的研究结果显示,位于miRNA种子区域的多态会影响miRNA的表达及其与靶基因的结合。例如,位于miR−125a种子区域的多态显著的抑制了pri/pre−miRNA的生成,导致miRNA表达减少。miR−206通过靶向ERα上的两个靶点调节其表达,而位于miR−206种子区域的多态导致两个靶点都失活,消除了其与原来靶基因的结合。理论上讲, miRNA种子区域的多态可以使得其调控的众多靶基因所行使的功能受到影响,但是这需要进一步的实验证实。

2. 位于miRNA的3' 容错区域(3' MTR) 与种子区域不同的是,3' MTR区域允许一定程度的碱基错配。但是,这一区域存在的插入、缺失或者移位的多态位点依然可能会对miRNA调控靶基因这一过程产生影响。具体的作用机制还有待进一步研究证实。

二、miRNA靶点的多态 》》》

一般情况下, miRNA调节基因表达的方式是与靶基因的3' UTR区域结合,进而降解靶

向的mRNA或者抑制这个mRNA的翻译过程。因此miRNA靶基因序列上的多态性可以干扰miRNA对靶基因的结合事件，而发生在miRNA靶基因序列上的功能性多态位点就被定义为miRNA靶点的多态。因为人类基因组中基因的3' UTR区域的序列保守性较差，所以与miRNA基因的多态性相比，miRNA靶位点具有更丰富的多态变异。全基因组关联分析（genome-wide association study，GWAS）发现，与编码区非同义突变相比，调控区域的变异更可能引起人类疾病。由于miRNA靶点多态可能干扰或破坏miRNA的结合位点或者产生一个新的miRNA的结合位点，进而引起靶基因调控的失调，所以它能影响多种生物学过程（例如，药物吸收通路、代谢、药物抵抗），并引发人类疾病（如乳腺癌、结肠癌、糖尿病以及心血管疾病等）。根据靶基因多态性位点与miRNA结合位点的位置关系，miRNA靶点的多态性可以分为miRNA结合位点上的多态性和miRNA结合位点周围的多态性。

（一）miRNA结合位点上的多态性

成熟miRNA长约22个碱基，miRNA 5' 末端的2–8个碱基被定义为miRNA的种子区域，经常与靶基因3' UTR的结合区域发生精确的互补匹配，发生在靶基因miRNA结合位点上的多态可以破坏miRNA对基因的调控作用（包括增强或者减弱miRNA与mRNA的结合）。例如在鳞状细胞癌（SCCHN）中SNP（rs8126 T—>C）落入到了基因TNFAIP2的3' UTR上。这个区域恰好是miR-184在其靶基因*TNFAIP2*上的种子结合区域，就使得这个SNP能够影响靶基因TNFAIP2的表达，改变鳞状细胞癌的易感性。*miR-582*在结肠癌中表达，而基因*CD86*上的SNP使得*miR-582*与其结合松散，因此导致*CD86*的表达升高，而且*CD86*可以激活炎性因子IL-4的表达，解释了*miR-582*结合位点上的多态位点rs17281995对结直肠癌风险的贡献。除此之外靶基因结合位点上的多态也可能会产生新的miRNA结合位点，进而导致miRNA对靶基因的调节发生紊乱，从而影响疾病的发生与发展。在12万个已知的发生在基因3' UTR上的SNP中，约有17%的多态会破坏推测的保守的或者不保守的miRNA结合位点，而且根据Patrocles数据库，有8.6%的多态会产生新的预测的miRNA靶位点。

（二）miRNA结合位点周围的多态性

miRNA靶点的多态除了位于miRNA与靶mRNA的结合位点，还可以位于miRNA结合位点周围的功能区域，这些多态也会影响miRNA对靶基因的调控作用。因为miRNA对靶mRNA发挥功能需要和一些辅助蛋白（比如AGO等）共同作用，形成RNA诱导的沉默复合物（RNA-induced silencing complex，RISC）以及RNA结合蛋白（RBPs）。此外，mRNA的3' UTR区域还存在着一些调控元件的结合位点，这些元件包括蛋白或蛋白复合物、细胞质多腺苷酸化元件（cytoplasmic polyadenylation elements，CPE）、六聚核苷酸AAUAAA等。例如mRNA的3' UTR上存在长约50个碱基的富含AU的序列（ARE），ARE与mRNA的稳定性密切相关，一些蛋白（比如Dicer1和Ago1）可以辅助ARE对mRNA的降解。研究发现miR-16可以与ARE互补，这说明miRNA可能会和ARE顺式调控元件相互作用，参与控制mRNA的降解。因此位于miRNA结合位点附近的多态会影响miRNA对靶基因的调控。Mishra发现*miR-24*结合位点下游位置上有一个多态性位点829（C<—>T），导致细胞对甲氨蝶呤的敏感性发生变化。正常生理状态下，*miR-24*可以结合到二氢叶酸还原酶（DHFR）的3' UTR上对该靶基因的表达起抑制作用，细胞表现为对甲氨蝶呤有较高的敏感性。但是当该结合位点下游第14个碱基发生突变

(T—>C),破坏了*miR-24*与DHFR的结合,无法降解DHFR的mRNA,导致靶基因二氢叶酸还原酶水平升高。另一方面,mRNA的3' UTR区域会形成一定的二级结构。Kedde M等人发现RNA结合蛋白Pumilio-1(PUM1)可以与基因*p27*的3' UTR互作,诱导RNA局部结构的改变,进而增强了miR-211和miR-222对*p27*的3' UTR的接近性,提高了miRNA对*p27*的有效抑制。研究表明多数的miRNA在 3' UTR上的结合位点都具有较简单的二级结构,这种简单的二级结构允许miRNA介导的RISC复合物接近结合位点并且发挥功能,所以位于miRNA靶点附近的多态位点可能引起mRNA 3' UTR二级结构的改变,从而影响miRNA与靶基因的结合。

三、miRNA合成机制中的单核苷酸多态

近些年的研究使得人们对于miRNA的合成机制有了更多的了解。miRNA的合成是一个复杂的生物学过程,主要分为三个阶段: DNA转录与加工合成pre-miRNA; pre-miRNA转运出核; 剪切合成成熟miRNA。miRNA合成过程中的每个阶段都有不同的蛋白质参与其中,这些与miRNA合成密切相关的蛋白质包括Drosha/DGCR8、核酸转运蛋白XPO5/RAN-GTP和Dicer/TRBP。因此,影响这些关键蛋白质表达的多态,会参与调节miRNA的合成及其生物学功能。本节主要讲述在miRNA的合成机制中,不同阶段相关蛋白质的多态对miRNA的影响。

Drosha和DGCR8加工DNA转录生成的pri-miRNA,形成约70个核酸长度的pre-miRNA。研究发现Drosha的低表达与癌症的预后和复发有密切关系,然而全基因组关联分析表明Drosha和DGCR8上的SNP并没有显示癌症易感性特征。XPO5和RAN-GTP介导了pre-miRNA从细胞核到胞质的过程,XPO5的异常表达会影响成熟miRNA的合成,进而影响了许多疾病的发生、发展。例如,研究发现XPO5在肺癌中频繁下调,而在高级前列腺癌样本中上调。由于XPO5和RAN参与miRNA的合成加工过程,所以这些基因上的SNP会影响miRNA的稳定性。Horikawa等人发现XPO5上的SNP(rs11070)增加了肾细胞癌的发病风险。Ryan等人对RAN的SNP(rs14035)研究结果表明,RAN上的多态能够阻断miRNA的核转运过程,导致miRNA合成受到抑制。这个SNP的祖先等位位于*miR-575*的绑定位点,突变体的形成又构建了*miR-182**的绑定位点,因此,RAN的多态还参与了miRNA对靶基因的调节过程。

Pre-miRNA转运出核之后,通过Dicer和TRBP的剪切作用,形成成熟的miRNA。研究发现低水平的Dicer与癌症患者的低存活率相关。Dicer的3' UTR上的SNP(rs3742330)能够增加黏膜白斑病和黏膜红斑病的恶化风险。另一项研究的结果发现Dicer的单体型与肾细胞癌的存活具有显著的关联,单体型AA和GA能显著增加肾细胞癌患者的死亡率。Melo的研究组识别了TRBP上的两个移码突变,这两个移码突变诱导产生了未成熟的终止密码子,进而使得TRBP表达下降。TRBP参与调节miRNA合成中Dicer的稳定性,当TRBP表达降低时,会导致Dicer不稳定和miRNA的合成减少。

RISC参与指导单链的成熟miRNA识别和靶mRNA的特定位点绑定。该复合物中GEMIN3上的非同义SNP(rs197412)的变异等位基因降低了口腔疾病的发病风险。GEMIN3的另一个SNP(rs197414)与膀胱癌和食管癌的高发病风险显著相关。因此,GEMIN3的变异能够影响miRNA的内稳态,进而调节细胞的信号通路。

最近,研究发现肿瘤抑制基因*p53*参与miRNA的合成过程。*p53*与*p68*、Drosha互作,促进了pri-miRNA到pre-miRNA加工过程。*p53*的变异与癌症之间存在密切关系。因此,*p53*上与

miRNA合成相关的突变或者SNP可能会增加或者降低癌症的发生风险。

四、miRNA多态改变表观遗传调控

miRNA的表达受表观遗传沉默的影响。miRNA的表观沉默最初是在乳腺癌的发病过程中被发现。在乳腺癌的早期,某些miRNA的表观沉默是很频繁的。很多miRNA都由于异常的高甲基化而受表观遗传沉默的影响。Lehmann等针对71例乳腺癌患者研究发现miR-9-1、miR-124a3、miR-148、miR-152和miR-663在34%~86%的患者中具有异常的高甲基化。因此,miRNA的异常高甲基化与癌症的发生密切相关。而在肺癌中Let-7家族是下调的,已有证据证明了Let-7a-3是低甲基化的,这也证明了miRNA可能在恶性肿瘤中具有两种作用。引起miRNA表观遗传调控改变的多态研究是一个新的研究领域。由于miRNA多态所导致的原癌基因或抑癌基因表观遗传调控的缺失或获得在细胞中可能具有决定性的影响,因此人们可以利用改变表观遗传调控的miRNA多态来研究疾病发生的机制。

虽然现在已经有case-control实验探索了异常的表观调控和癌症发病风险之间的关系,但是这种遗传变异形式所蕴含的意义在整体水平上还是未知的。此外,CpG岛中的miR-SNP很可能影响miRNA的表达模式,进而影响癌症的易感性。一个miRNA的启动子中出现了一个SNP(无论是不是CpG岛),它可能会影响miRNA的表达水平。确实Sevignani等人发现易感肿瘤的大鼠中大多数是miRNA基因序列上的差异,而不是大鼠的抗肿瘤基因的启动子上的改变。

【例10-5】结合miRNA表达谱和SNP谱分析miRNA与疾病的关系

在理解了SNP与miRNA之间的关系后,我们通过一个基本的例子,简要介绍一下miRNA数据是如何与SNP数据结合使用的。这个例子的目的,就是通过加入SNP信息,优化差异表达miRNA的计算方法。

1. 数据 首先,我们需要获得有类标签(疾病/正常)的miRNA表达谱和SNP谱,同时还需准备miRNA在染色体上的位置信息,以及SNP在染色体上的位置信息。

2. 检验 在这一步骤中,我们需要使用不同方法对miRNA和SNP数据进行分析,确定某个miRNA或SNP是否与疾病相关。与传统的差异表达基因计算方法类似,我们使用两样本t检验对miRNA数据进行处理。公式如下:

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} - \frac{\sigma_2^2}{n_2}}} \qquad (10\text{-}6)$$

在面对海量数据时,由于检验次数过多,t检验这种传统的单变量分析方法的检验效能是较低的。因此,我们可以加入SNP数据进行辅助分析。对于离散型的SNP数据,我们使用卡方检验对数据进行处理:

$$K_1 = \sqrt{\frac{\sum_{i=1}^{2} S_i}{\sum_{i=1}^{2} R_i}} \qquad K_2 = 1/K_1 \qquad (10\text{-}7)$$

$$\chi^2 = \sum_{i=1}^{2} \frac{\left(K_1 R_i + K_2 S_i\right)^2}{R_i + S_i} \qquad (10\text{-}8)$$

其中，R和S分别代表正常组和疾病组的样本计数，而i则代表SNP亚型的分类数目。公式所描述的是当基因型数量为2的情况。例如，R_1表示正常样本中基因型为“CC”纯合的样本数目，而S_2则表示疾病样本中基因型为“CG”杂合的样本数目。

3. 整合 在分别得到所有miRNA与SNP的显著性之后，如何将它们进行统一就成了首要的问题。为此，我们需要将有联系的miRNA与SNP进行多对多映射。通过前面的内容我们知道SNP可以通过很多途径影响miRNA的表达，为了便于理解，在这个例子中我们仅考虑落在miRNA内的SNP对miRNA表达的影响。通过miRNA和SNP在染色体上的位置信息，我们就可以找出所有落入miRNA内的SNP位点了。当许多SNP落入同一个miRNA时，我们仅取x^2值最大的SNP，这样就使得miRNA与SNP一一对应了。

此时，我们使用meta分析中常用的Fisher组合概率检验，就可以将对miRNA和SNP检验得到的两个p值整合成一个统计量了。公式如下：

$$\chi^2 = -2\sum_{i=1}^{k} \ln\left(p_i\right) \qquad (10\text{-}9)$$

其中，k表示检验个数，即k=2。这样，对这个χ^2统计量进行检验就得到了最终的p值，描述了在SNP信息辅助的情况下，我们得到的两种状态下（疾病/正常）miRNA表达没有变化的概率。当它显著时，就可以说明miRNA在两种状态下表达不同了。

（肖 云 许超汉 平艳艳 徐 娟 李 霞）

参考文献

1. Li X, Jiang W, Li W, et al. Dissection of human miRNA regulatory influence to subpathway. *Brief Bioinform* 2012, 13(2): 175–186.
2. Li X, Wang Q, Zheng Y, et al. Prioritizing human cancer microRNAs based on genes' functional consistency between microRNA and cancer. *Nucleic Acids Res* 2011, 39(22): e153.
3. Xu J, Li CX, Li YS, et al. MiRNA–miRNA synergistic network: construction via co-regulating functional modules and disease miRNA topological features. *Nucleic Acids Res* 2011, 39(3): 825–836.
4. Xiao Y, Xu C, Guan J, et al. Discovering Dysfunction of Multiple MicroRNAs Cooperation in Disease by a Conserved MicroRNA Co-Expression Network. *PLoS One* 2012, 7(2): e32201.
5. Jima DD, Zhang J, Jacobs C, et al. Deep sequencing of the small RNA transcriptome of normal and malignant human B cells identifies hundreds of novel microRNAs. *Blood* 2010, 116(23): e118–127.
6. Creighton CJ, Reid JG, Gunaratne PH. Expression profiling of microRNAs by deep sequencing. *Brief Bioinform* 2009, 10(5): 490–497.
7. Friedlander MR, Chen W, Adamidi C, et al. Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* 2008, 26(4): 407–415.
8. Winter J, Jung S, Keller S, et al. Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nat Cell Biol* 2009, 11(3): 228–234.

9. Krol J, Loedige I, Filipowicz W. The widespread regulation of microRNA biogenesis, function and decay. *Nat Rev Genet* 2010, 11(9): 597–610.

10. van Kouwenhove M, Kedde M, Agami R. MicroRNA regulation by RNA-binding proteins and its implications for cancer. *Nat Rev Cancer* 2011, 11(9): 644–656.

11. Avraham R, Yarden Y. Feedback regulation of EGFR signalling: decision making by early and delayed loops. *Nat Rev Mol Cell Biol* 2011, 12(2): 104–117.

12. Mestdagh P, Lefever S, Pattyn F, *et al*. The microRNA body map: dissecting microRNA function through integrative genomics. *Nucleic Acids Res* 2011, 39(20): e136.

13. Joung JG, Hwang KB, Nam JW, et al. Discovery of microRNA–mRNA modules via population-based probabilistic learning. *Bioinformatics* 2007, 23(9): 1141–1147.

14. Joung JG, Fei Z. Identification of microRNA regulatory modules in Arabidopsis via a probabilistic graphical model. *Bioinformatics* 2009, 25(3): 387–393.

15. Croce CM. Causes and consequences of microRNA dysregulation in cancer. *Nat Rev Genet* 2009, 10(10): 704–714.

16. Marson A, Levine SS, Cole MF, et al. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* 2008, 134(3): 521–533.

17. Su X, Chakravarti D, Cho MS, et al. TAp63 suppresses metastasis through coordinate regulation of Dicer and miRNAs. *Nature* 2010, 467(7318): 986–990.

18. Martinez NJ, Walhout AJ. The interplay between transcription factors and microRNAs in genome-scale regulatory networks. *Bioessays* 2009, 31(4): 435–445.

19. Glinsky GV. An SNP-guided microRNA map of fifteen common human disorders identifies a consensus disease phenocode aiming at principal components of the nuclear import pathway. *Cell Cycle* 2008, 7(16): 2570–2583.

20. Kim J, Bartel DP. Allelic imbalance sequencing reveals that single-nucleotide polymorphisms frequently alter microRNA-directed repression. *Nat Biotechnol* 2009, 27(5): 472–477.

第十一章

CHAPTER 11

计算表观遗传学

COMPUTATIONAL EPIGENETICS

第一节

引　　言

Section 1　Introduction

在证实DNA承载着可遗传的分子信息之前，科学家已经发现尽管机体的所有细胞拥有相同的遗传信息即相同的基因组DNA序列，却并不是每个基因在各个细胞内都具有表达活性。Waddington 把这一现象定义为"表观遗传学"（epigenetics），这门学科主要是探索不涉及DNA序列改变，由DNA甲基化谱、染色质结构状态等改变而导致基因功能的变化并在细胞代间遗传的现象的本质和规律。

表观遗传学研究已有60多年的历史，近些年来随机的或环境诱导的表观遗传改变已经成为生命科学及现代医学研究领域的热点，不仅在癌症中发现了表观遗传修饰的改变，在其他非癌症的疾病包括免疫系统疾病、心血管疾病、神经性疾病和代谢性疾病等的发病机制的研究中也发现与表观遗传异常有关。

高通量实验技术的发展及其在表观遗传学研究领域的应用，已经从基因组水平检测出一些导致疾病发生的表观遗传异常，包括基因组局部或全局的DNA甲基化的改变和染色质蛋白质修饰的错误发生、分布或功能异常引起基因表达失调。因此表观遗传学作为媒介架起了遗传学和环境之间的桥梁，通过对表观遗传现象和机制的深入研究有助于理解个体间遗传背景、环境及衰老与疾病之间的关系。

计算表观遗传学的研究浪潮源于高通量实验技术下飞速出现的海量基因组范围的表观遗传修饰的数据，生物信息学的算法和工具，对解决表观遗传学领域的各种问题起到了重要作用。结合传统的基因组学、计算机科学、数学以及生物化学、蛋白质组学所获得的表观遗传学的结论，不仅可以指导实验设计，还能实现仅仅由传统实验方法不能做到的详细分析复杂的基因组信息的目的。基于计算表观遗传学发展起来的方法和工具，大规模的分析不依赖于基因序列的可遗传的显型改变，基因功能和基因表达，为了解转录调控、发育和疾病过程提供了高效实用的工具。

第二节

表观基因组图谱绘制

Section 2 Mapping of Epigenome Profiles

表观遗传学是当前生物医学领域中发展较快的研究热点之一,已经在人类的多种疾病中发现了表观遗传改变现象。在过去的10年内,随着生物实验技术的革命,可以从全基因组水平考察表观遗传修饰的变化,被称为"表观基因组"研究。表观基因组是由所有基因组范围的染色质修饰组成,包括DNA甲基化和组蛋白修饰。染色质修饰的不稳定性使得表观基因组呈现出动态变化,为生物体提供了响应和适应环境信号的基因表达调控的机制。芯片技术及新一代测序技术为研究者提供了绘制各种生命状态(如疾病状态的组织或细胞和正常组织或细胞)全基因组范围内高分辨率的DNA甲基化和蛋白质翻译后修饰(如组蛋白修饰)图谱的工具,实现从单个基因到基因组全局水平的研究人类疾病的发生、发展过程。

一、绘制基因组范围的DNA甲基化谱

(一) DNA甲基化

DNA甲基化是目前为止研究比较成熟的表观遗传修饰之一,是导致一些人类疾病发生发展的重要表观遗传修饰改变,特别是在肿瘤等疾病的发病中。哺乳动物中DNA甲基化是通过DNA甲基转移酶(DNA methyltransferase)的作用,在5′-胞嘧啶-鸟嘌呤-3′二核苷酸(CpG)内的胞嘧啶第5位碳原子上添加来自于S-腺苷甲硫氨酸(SAM)的甲基(CH_3)基团,目前的研究发现在非CpG的胞嘧啶上也可能发生甲基化(图11-1)。DNA甲基化通常通过影响甲基化-敏感的DNA结合蛋白和(或)改变DNA到启动子的接近性组蛋白不同修饰的相互作用而引起基因沉默。大量研究已经表明,DNA甲基化的功能是多种多样的,包括使转录原件沉默,对发育相关基因的调控和转录噪声的减少。研究已经发现异常的甲基化模式发生在多种人类疾病中,包括癌症,ICF综合征(immunodeficiency-centromeric instability-facial anomalies syndrome,表现为免疫缺陷、着丝粒不稳定性、面部异常),ATRX综合征(α-地中海贫血,表现为精神发育迟缓),以及脆性X染色体综合征等。

(二) DNA甲基化的分布和检测

1. 基因组上DNA甲基化检测技术发展　测定全基因组范围的DNA甲基化对于理解表

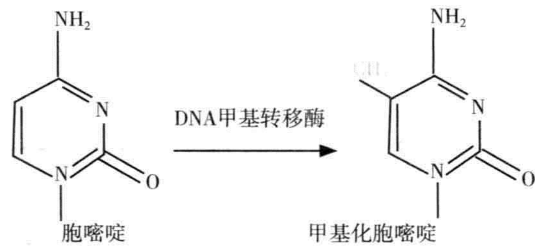


图11-1 DNA甲基化的发生

观遗传学的作用机制是极其重要的。早在1970年就开始了通过甲基化敏感性限制内切酶的方法测定全局的DNA甲基化含量，如Kawai等开发了限制性内切酶或者限制性的标记的基因组扫描(restriction landmark genome scanning, RLGS)方法测定少量基因区域在各组织间的DNA甲基化。然而，这种方法实验精度和广度受到酶切位点的限制，不适用于完全的基因组扫描。

为了解决这个问题，Frommer等人在1992年引入亚硫酸氢盐转换技术来精确地测定DNA甲基化，使甲基化测定技术取得了革命性进展，该技术被誉为测定胞嘧啶甲基化的"金标准"。重亚硫酸盐预处理方法的发现引发了甲基化胞嘧啶高精确性测定的革命，这种方法已被多个大型甲基化测定计划所采用，其中就包括人类表观基因组计划(HEP)。然而，由于该方法实验成本较高，限制了其在基因组范围的应用。

为了解决这些限制，Weber等人基于染色质免疫共沉淀原理，利用特定抗体对甲基化区域的亲和纯化作用，开发了甲基化DNA免疫共沉淀测定技术(methylated DNA immunoprecipitation, MeDIP)，该技术采用甲基化胞嘧啶特异的抗体获得甲基化的DNA序列片段。MeDIP与寡核苷酸芯片的结合(MeDIP-chip)为DNA甲基化谱的测定提供了一个有效的手段。

另外，下一代测序技术的发展促成了甲基化测定技术的第二次革命，开发了一些基于测序的甲基化实验技术，如MethylC-Seq、RRBS和MeDIP-seq等，这些技术中的大部分方法都可以实现测定全基因组范围内单碱基水平的DNA甲基化数据，已经被广泛用于测定大型基因组区域中的DNA甲基化模式。

目前已有的测定DNA甲基化的技术(表11-1)产生的数据大部分都可以通过0到1(或0%到100%)之间的连续值来表示甲基化程度的高低。因此高精度的甲基化数据使得定量解释DNA甲基化差异调控基因表达的机制成为可能。DNA甲基化模式的描述及广泛的DNA甲基化谱绘制可以帮助理解在发育的特定阶段以及疾病的发生中基因组的表观遗传改变如何使特定基因表达模式发生变化。

表11-1 各种测定DNA甲基化的技术

下一代测序技术			芯片杂交技术			其他
酶切	亲和纯化	重亚硫酸盐	酶切	亲和纯化	重亚硫酸盐	
⊙MSCC	⊙MeDIP-seq	⊙MethylC-Seq	⊙CHARM	⊙MeDIP	⊙GoldenGate	⊙MS-AP-PCR
⊙Methyl-MAPS	⊙MIRA-seq	⊙RRBS	⊙MCAM	⊙MIRA	⊙Infinium	⊙RLGS
⊙Methyl-seq	⊙MiGS	⊙BC-seq	⊙DMH	⊙mDIP	⊙BiMP	⊙Sanger BS

续表

下一代测序技术			芯片杂交技术			其他
酶切	亲和纯化	重亚硫酸盐	酶切	亲和纯化	重亚硫酸盐	
		⊙BSPP	⊙MMASS	⊙mCIP		⊙MS-SNuPE
			⊙HELP			⊙COBRA
			⊙MethylScope			⊙Southern blot
						⊙AIMS

注：数据来源：Laird，PW（2010）Nat Rev Genet，11：191-203。

2. DNA甲基化检测技术介绍

（1）核酸内切酶消化：在分子生物学研究领域，限制性核酸内切酶是一种有力的研究工具。每个序列特异限制酶都对应一个DNA甲基化转移酶，该酶能保护内源性DNA免受外界酶的影响，一些限制性内切酶在甲基化的胞嘧啶处结合的明显减少，因此这些酶切割的模式能够用来判定DNA序列的甲基化状态。其中使用最广泛的甲基化敏感限制酶对包括HpaII-MspI（CCGG）和SmaI-XmaI（CCCGGG）。

以5'-CCGG-3'位点为例，不管其中的CpG甲基化与否，MsqI均能切割CCGG序列，而HpaII只切割没有甲基化的CCGG序列。因此，如果两个酶消化的片段相同，说明该CpG位点是未甲基化的；若不同，则表明该CpG位点是甲基化的，从而利用该限制酶对可以用来区分CCmeGG和CCGG。

基因组序列的酶切处理与不同的碱基测定技术结合产生了不同的DNA甲基化测定技术。20世纪70年代和80年代初期，甲基化敏感限制酶的消化DNA序列，结合凝胶电泳和Southern blots杂交，用于一些基因座特异的研究。这种酶切消化甲基化敏感位点与PCR技术的结合，是一种非常敏感的技术，至今仍应用于一些研究中。

从20世纪90年代起，人们又开发了多种基于酶切的方法测定基因组范围的DNA甲基化。其中，限定标记基因组扫描法（RLGS）是第一个用于检测基因组范围DNA甲基化谱的技术，该方法是基于二维凝胶电泳来测定实验样本和对照样本间的甲基化差异，已经广泛应用于筛选癌症特异的印记基因/位点。类似的方法还有甲基化敏感随机性引物PCR（methylation-sensitive arbitrarily primed PCR，MS-AP-PCR）和甲基化间区位点扩增（amplification of inter-methylated sites，AIMS）等。

目前，随着芯片技术以及测序技术的不断发展，这些基于凝胶电泳的DNA甲基化技术的应用越来减少。甲基化CpG岛扩增法（methylated CpG island amplification，MCA）就是酶切方法与芯片技术结合的典型技术之一，该方法利用的是甲基化敏感限制酶对SmaI-XmaI，两种酶对SmaI和XmaI对CCCGGG片段甲基化敏感性和切割方式存在差异性，无论该片段中的CpG位点是否甲基化，XmaI均对其进行切割，而SmaI只切割未发生CpG甲基化的片段。但是，该方法与其他基于4bp碱基识别序列的其他酶切方法相比，MCA方法的精度较低。一种替代方法是差异甲基化杂交（DMH），该方法基于双色微阵列分别与甲基化敏感限制酶消化和模拟消化DNA序列杂交，再根据相对的荧光信号强度来提取阵列上对应位点的DNA甲基化信息。基于DMH，人们提出了很多改进的方法，如利用内切核酸酶McrBC的方法，与甲基

化敏感酶相比，该方法能够为高密度甲基化的区域中甲基化水平的测定提供更高的精度。其中最优化的方法是全面高通量芯片相对甲基化技术（comprehensive high-throughput arrays for relative methylation，CHARM），该技术已经被广泛用于研究癌症特异的以及干细胞特异的DNA甲基化区域分析。

逐渐下一代测序技术也用于测定基于酶切方法富集出的甲基化片段。基于序列的分析更加灵活和强大，因为它允许等位基因特异DNA甲基化分析，不需要预先设计探针，即使在较少的DNA样品中仍能覆盖全基因组。其中最为常用的技术是甲基化测序（Methyl-seq）和甲基化敏感切割位点计数（methylation-sensitive cut counting，MsCC）等。

（2）亲和性富集：目前的研究已经证实染色质免疫共沉淀（ChIP）是一种对于组蛋白修饰全基因组研究特别有用的技术。相似的，甲基化胞嘧啶特异的抗体（在变性DNA附近）或者用对甲基化的局部基因组DNA 有亲和力的甲基绑定蛋白，可以用来测定基因组范围的DNA甲基化。

Cross等人利用甲基绑定蛋白MECP2，第一次实现了甲基化DNA的亲和纯化。亲和纯化后，将甲基化相关的片段杂交到芯片来测定DNA甲基化的水平，这种方法被命名为MeDIP-chip。除了与芯片杂交的结合外，目前人们更多地将亲和纯化与下一代测序技术联合起来，称这种方法为MeDIP-seq。

基于亲和纯化的DNA甲基化测定技术已经广泛地应用到检测植物、小鼠以及人类等各种细胞的甲基化数据谱中。尽管这种方法能够快速有效地对基因组范围的DNA甲基化进行评估，然而这些甲基化信息的精度仅限于基因组区域，并不是单碱基水平的，而且对于不同CpG密度的基因组区域测定的DNA甲基化信息，还需要进行实验的或者生物信息学的校正处理。

（3）重亚硫酸盐转换：20世纪90年代，人们发现通过亚硫酸氢钠处理，非甲基化的胞嘧啶（C）被脱氨基作用而变成尿嘧啶（U），在随后的PCR反应中尿嘧啶（U）变成胸腺嘧啶（T）；而甲基化的胞嘧啶（5-MeC）不能被脱氨基，这一发现促成了DNA甲基化分析领域的革命性进展。亚硫酸氢钠将非甲基化胞嘧啶转换为胸腺嘧啶的反应使得许多新的DNA甲基化检测和分析技术的开发成为可能。

最初的重亚硫酸盐转换DNA的分析是由单位点克隆PCR产物的桑格测序实现的。许多增强的功能从那以后也有所发展，包括PCR产物的定量直接桑格测序以及更为自动化的DNA甲基化测定技术。

人们将重亚硫酸盐处理和芯片杂交技术相结合，开发出了重亚硫酸盐甲基化谱（bisulfite methylation profiling，BiMP），该方法通过杂交阵列来分析重亚硫酸盐处理过的DNA，需要在对一个专用的寡核苷酸阵列杂交前，基因组单个区域扩增，实质上是适当的比例增大位点特异阵列。值得注意的是，这种方法依赖于非重亚硫酸盐转换DNA建立的微阵列，因此，作为非甲基化胞嘧啶残留物转换引起的错配的结果，潜在的甲基化靶序列内外全部杂交信号都是低的。由于它们保留更多的Cs，产生出相对强的信号，密集胞嘧啶甲基化区域受到的影响最少。所以，BiMP仅适用于小基因组甲基化密集区。

此外，Illumina公司将其GoldenGate 磁珠技术进行了改进，并与重亚硫酸盐处理相结合，用其来查询人类基因组DNA样本中CpG位点的DNA甲基化。该技术首先通过重亚硫酸盐对DNA序列进行处理，然后用甲基化和非甲基化特异的引物提取甲基化和非甲基化的CpG位点所在的片段，并用不同的荧光染料标记，随后将提取出的产物结合到磁珠芯片上进行测

定，最后根据甲基化和非甲基化的磁珠的计数来估计每个位点的甲基化水平。该技术最初只能测定1536个不同CpG位点，随后将位点数目扩展至27 578个。目前该方法已经支持同时测定>485 000个CpG点，全面覆盖了96%的CpG岛，并根据需求加入了CpG岛以外的CpG位点、人类干细胞非CpG甲基化位点、正常组织与肿瘤（多种癌症）组织差异甲基化位点、编码区以外的CpG岛、miRNA启动子区域和已通过GWAS的疾病相关区域的位点，每张芯片可平行进行12个样本的检测，因此它非常适合大量样本分析。

尽管重亚硫酸盐转换与DNA到阵列杂交能够提供多样本DNA甲基化的分析，但是显然这种方法测定的位点数量仍然有限。人们基于下一代测序技术，开始了多种高通量单碱基的DNA甲基化测定技术。由于哺乳动物庞大的基因组和复杂的细胞状态，目前基于PCR或全基因组鸟枪法效率都很低。

Meissner等开发了简约重亚硫酸盐测序技术（reduced representation bisulphite sequencing, RRBS），该方法仅对通过BglII 或者MspI从庞大的基因组选择特定的区域进行测序，提高了基因组范围内CpG位点甲基化状态测定的效率。在另一个重亚硫酸盐测序技术（bisulphite conversion followed by capture and sequencing，BC-seq）中，则通过芯片获取重亚硫酸盐处理后DNA并用PCR扩增，从而为创建测序文库获得充足的DNA。最终的综合单碱基分辨率DNA甲基化分析的技术则是全基因组重亚硫酸盐测序。全基因组鸟枪重亚硫酸盐测序（whole-genome shotgun bisulfate sequencing, WGSBS）在Illumina 基因组分析平台上已经得以实现，并对小真核生物基因组（如拟南芥）和哺乳动物DNA（如人类）进行的测定分析。虽然在哺乳动物基因组上约十分之一的CpG二核苷酸仍然难以被亚硫酸氢盐转换的片段覆盖，但增长的读取片段的长度和双末端测序策略促成了WGSBS的实现及其更为广泛的应用。

（三）不同甲基化谱方法间的比较

DNA甲基化方法的直接比较受不同技术的复杂性和差异限制。许多方法都有竞争优势和劣势。对于DNA甲基化测定技术的选择除了考虑覆盖率和分辨率，还受样本数量和DNA质量和数量的影响。此外，还要考虑被研究的物种，如基于芯片的方法，需要已有的物种芯片的支持；而基于测序的方法，则由于参考基因组的存在，一般可应用于任意物种。

二、高通量染色质修饰谱的测定

（一）组蛋白修饰

组蛋白修饰（histone modification）是真核生物中染色质的主要修饰之一，具有组织特异性，对外界环境变化敏感，并对基因表达起到关键调控作用。在疾病细胞中，组蛋白修饰模式亦发生改变。尽管发现组蛋白修饰几十年，但对它的知识积累的快速增加却是在最近几年。随着高通量实验技术的推广，绘制的基因组范围的组蛋白修饰图谱不仅增进了人们对组蛋白修饰模式的认识，也有助于理解组蛋白修饰在疾病发生过程所起的作用。

在早期的染色质研究中，染色质被描述成“一串线上的珠子”，这些珠子就是核小体。每个核小体由八个核心组蛋白（包括H3，H4，H2A和H2B各两个）及缠绕在八聚体组蛋白表面的DNA序列构成。核心组蛋白通过H1组蛋白相互连接（图11-2）。

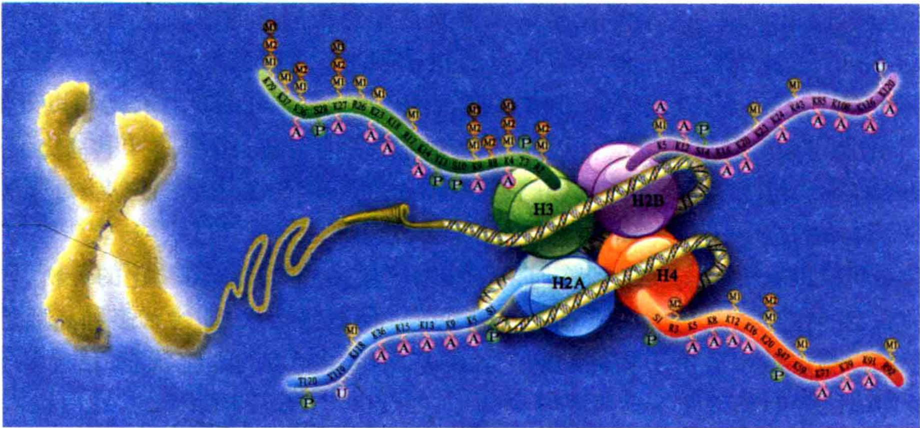


图11-2 组蛋白修饰

来源于：Zhang Y.HHMD：the human histone modification database.Nucleic Acids Res.2010；38：149-154.

在核心组蛋白的末端有组蛋白尾巴，易受多种共价修饰，包括赖氨酸和精氨酸的甲基化，赖氨酸的乙酰化，丝氨酸的磷酸化（表11-2）。组蛋白修饰通过降低和DNA结合的亲和力，还通过征调更多的染色质重构复合物来影响核小体组装成更高维的包装结构。这些修饰的不同组合模式内存储的潜在信息形成"组蛋白密码"的假说，它们的特定组合组合表明基因座特定的转录模式。

当编码组蛋白修饰基因的改变和组蛋白修饰模式的紊乱与疾病的发生有密切的关系。例如，多硫蛋白EZH2，组蛋白H3赖氨酸37（H3K27）甲基转移酶在前列腺癌中过表达，而H4K16乙酰化和H4K20三甲基化（H4K20me3）在淋巴癌和结肠癌中观察到全局的缺失。

表11-2 高通量实验测定的组蛋白修饰类型

组蛋白类型	组蛋白修饰
H2A	H2AK5ac，H2AK9ac，H2AZ
H2B	H2BK120ac，H2BK12ac，H2BK20ac，H2BK5ac，H2BK5me1，UbH2B*
H3	H3K14ac，H3K18ac，H3K23ac，H3K27ac，H3K27me1，H3K27me2，H3K27me3，H3K36ac，H3K36me1，H3K36me3，H3K4ac，H3K4me1，H3K4me2，H3K4me3，H3K79me1，H3K79me2，H3K79me3，H3K9ac，H3K9me1，H3K9me2，H3K9me3，H3R2me1，H3R2me2，H3ac*
H4	H4K12ac，H4K16ac，H4K20me1，H4K20me3，H4K5ac，H4K8ac，H4K91ac，H4Kac，H4R3me2，H4ac*

注：*没有使用特异的抗体。数据来源：Zhang Y.HHMD：the human histone modification database. Nucleic Acids Res.2010；38：149-154。

（二）组蛋白修饰的高通量谱绘制

目前测定组蛋白修饰的各种技术均依赖于染色质免疫沉淀（ChIP）技术。这项技术采用特定抗体来富集存在组蛋白修饰或者转录调控的DNA片段，通过多种下游检测技术（定量PCR，芯片，测序等）来检测此富集片段的DNA序列，已经广泛用于多个领域的染色质相关蛋白的研究（如组蛋白及其异构体，转录因子等），特别适用于已知启动子序列或整个基因位点的组蛋白修饰分析研究。当前基于ChIP测定组蛋白修饰的实验技术主要分为两类：ChIP与

芯片技术结合的ChIP-chip技术，ChIP与测序技术结合的ChIP-seq技术。

1. ChIP-chip 描绘组蛋白修饰

（1）ChIP-chip 原理介绍：染色质免疫共沉淀-芯片（ChIP-chip）的基本原理是在特定的实验条件下通过甲醛将组蛋白和DNA交联，并利用超声波将其打碎为一定长度范围内的染色体片段（<1kbp），然后通过组蛋白修饰特异性抗体沉淀复合物片段，从而特异性地提取组蛋白修饰修饰结合的DNA片段，并对这些片段的进行纯化，最后利用高通量芯片技术对片段进行检测，从而获得组蛋白修饰与DNA相互作用的信息。

ChIP与基因芯片相结合建立的ChIP-chip方法（图11-3A）已广泛用于特定组蛋白修饰的高通量筛选，从而高通量的筛选特定组蛋白修饰的靶向基因组DNA序列。之前的研究表明组蛋白修饰与基因组DNA结合能够调控染色质重塑和基因转录，因此对基因组范围组蛋白修饰分布的研究能够揭示疾病等过程中的表观遗传调控机制。

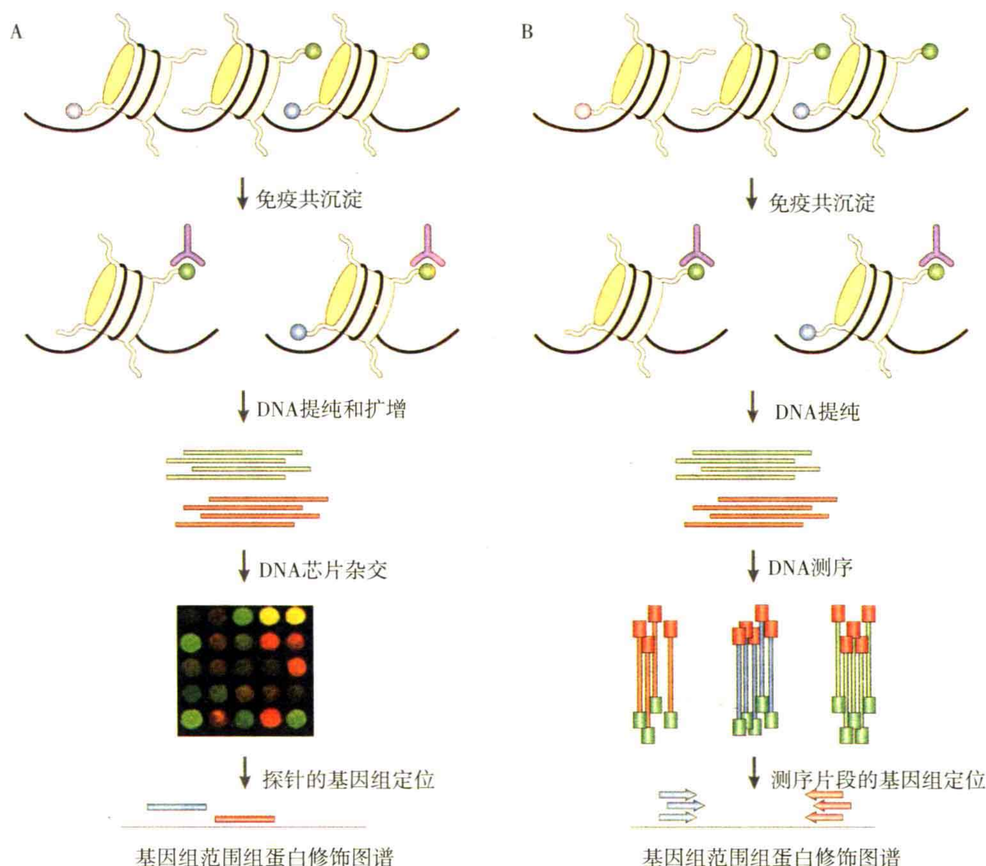


图11-3 ChIP-chip和ChIP-seq方法示意图

A. ChIP-chip技术流程；B. ChIP-seq技术流程

来源于：Schones DE.Genome-Wide approaches to studying Chromatin modifications,Nat Rev Genet. 2008；9(3)：179-191.

目前已经有研究者将该技术应用于测定疾病与正常样本间的差异组蛋白修饰模式，发现组蛋白修饰介导疾病中的表观遗传调控的异常。染色质免疫共沉淀技术与芯片技术相结合有助于科学家发明疾病的有效治疗方法。

（2）ChIP-chip技术测定组蛋白修饰：ChIP-chip技术在测定组蛋白修饰方面具有几个优点：①可以在细胞内进行组蛋白和DNA的交联反应；②能够得到各种特定待检细胞中组蛋白修饰与DNA的结合位置信息；③特定的组蛋白修饰抗体特异性地靶向待检修饰的相关位点；④能够进行全基因组范围组蛋白修饰的测定。

总之，ChIP-chip技术的发展为分析各种疾病状态及其对照组织中DNA与组蛋白修饰的相互关系提供了一个极为有力的工具。除了在测定组蛋白修饰方面的应用外，ChIP-chip在研究转录因子调控基因表达、增强子和隔离子等远端调控原件的测定以及染色体重塑中的应用也十分广泛。

（3）ChIP-chip技术在组蛋白修饰检测的应用：加州大学路德维格癌症研究所（ludwig institute for cancer research，LICR）的研究者利用ChIP-chip技术对人类细胞的物种核心的组蛋白修饰模式进行了测定，这五种修饰包括：H3ac、H4ac、H3K4me1、H3K4me2、H3K4me3，发现了人类基因组中重要的功能性元件。人类基因组是包裹在染色质当中的，确切地说是由组蛋白包裹DNA。针对全基因组序列，研究者分析了人类细胞的五种修饰与启动子和增强子的关系，发现已知的启动子、增强子附近被特有的组蛋白修饰标记，如H3K4me1和人类增强子相关。

根据这些特征，研究者开发了算法，识别出了几百个新的潜在增强子，这些都可能是具有潜在调控功能的基因组区域。该研究负责人Ren表示：这种方法的理论具有普遍性，而且可以运用这套相对公正的方法探索基因表达在患病情况下是如何变化的分子机制；这个方法的魅力所在是它依赖于组蛋白的化学特征，而不是DNA的；对于组蛋白修饰特征的解析将让科学家快速识别出基因的增强子和启动子，在此基础上可以进一步方便快捷地识别调控基因表达的因子；这种方法还可以用来识别在癌症发生过程中基因网络异常的发生，这将推动癌症检测技术的开发。

2. ChIP-seq 检测组蛋白修饰

（1）ChIP-seq的原理介绍及检测组蛋白修饰：染色质免疫共沉淀-测序（chromatin immunoprecipitation sequencing，ChIP-seq）另一种测定组蛋白修饰的高通量技术，与ChIP-chip不同的是，ChIP-seq通过对使用免疫共沉淀（ChIP）后对产生的DNA片段进行测序来获取组蛋白修饰与DNA序列的结合关系。同样地，组蛋白修饰特异的抗体对ChIP-seq的成功至关重要（图11-3B）。由于实验技术的不同，ChIP-seq的分析与ChIP-chip也有一定的差别。

ChIP-seq的第一步分析是将测序的DNA片段匹配到参考基因组上，目前已经有许多有效的生物信息算法用于短读物匹配。在读物匹配到参考基因组后，下一步的分析则是试图探测组蛋白修饰高度富集的基因组位置，这个过程被称作峰值探测（peak calling）。

目前广泛应用的有两种峰值探测方法。其中一个是基于固定长度窗口的方法，该方法使用一个从随即读取片段位置得到的经验背景模型用于估计显著性。然而，组蛋白修饰的区域的跨度比较大，有时是几百个碱基，而有时则跨越上千碱基，如抑制性的标记（H3K27me3和H3K9me3）发生在较长的基因组区域中，这样固定长度的窗口法不适用于探测区间可变的峰值区域。

另一种方法则基于隐马尔科夫模型，可得到可变长度的窗口，提高了峰值探测的精度。基于组蛋白修饰在特定细胞系中峰值的分布，人们可以研究组蛋白修饰与各种基因组调控原件的位置关系，从而揭示组蛋白修饰在转录调控中的作用。

（2）ChIP-seq 的应用：Zhao等人利用ChIP-seq构建人类CD4细胞中39种组蛋白修饰内的全基因组范围图谱，通过对这些组蛋白修饰及其调控原件如启动子，隔离子，增强子和转录区域的分析发现几个重要的结论：H3K27，H3K9，H4K20，H3K79和H2BK5的单甲基化都与基因活化有关；H3K79单甲基化与基因抑制有关；CTCF标示组蛋白甲基化区域的边界；染色体带型与特定的组蛋白修饰相关；T细胞相关的癌症中的染色体断列位点与H3K4相关的染色体区域有关；组蛋白乙酰化是一种关键的翻译后修饰模式，对组蛋白及其他蛋白的修饰都具有重要的作用，主要对转录的调节具有重要的意义。该研究的负责人表示：HDACs不仅仅具有抑制转录的功能还能修饰染色质中的活性基因；失活的基因首先通过与MLL介导的H3K4甲基化作用，随后接受HAT/HDAC的动态乙酰化和去乙酰化的作用，在阻止基因与PolⅡ结合而抑制基因表达的同时，可以使这些基因保持能够被激活的状态；一旦编码HATs和HDACs的基因发生突变，将导致多种疾病的发生，包括癌症。

3. ChIP-chip和ChIP-seq技术的比较　尽管ChIP-chip和ChIP-seq两种测定组蛋白修饰的技术都基于染色质免疫共沉淀，然而，由于二者采用的后续测定技术的不同，它们在分辨率、定量性、覆盖范围以及实验费用方面的差异较大（表11-3）。

表11-3　ChIP-chip和ChIP-Seq的比较

检测技术	ChIP-chip	ChIP-Seq
分辨率	30~100bp	1bp
分辨率的影响因素	探针密度	测序深度
定量性	受杂交效率影响	定量
动态量程	弱信号会被丢弃；强信号会饱和	无限制
覆盖范围	受芯片容量限制，局限于预设的基因组区域	可覆盖大部分基因组区域
全基因组范围实验费用	多	少
需要的DNA量	高	低
缺点	探针和非特异性区域杂交	测序数据受GC含量的影响

由于芯片技术中探针密度以及染色质长度的限制，ChIP-chip目前的分辨率仅为30~100bp，这样ChIP-chip测定的组蛋白修饰数据仅能反映较长区域的组蛋白修饰状态；而基于测序技术的ChIP-Seq则不依赖于芯片的限制，仅依赖于测序的深度，只要测序深度达到一定的量，就能够测定任何有组蛋白修饰靶定的基因组区域，因此ChIP-Seq的最高精度可以达到1bp，这可能也是目前ChIP-seq逐渐代替ChIP-chip成为主要的组蛋白修饰测定技术的主要原因之一。

除了在分辨率方面的差异，ChIP-chip和ChIP-seq另一个主要的差别在二者对组蛋白修饰进行定量性，ChIP-chip的定量性受到芯片技术过程中杂交效率的影响，且还可能发生弱信号被丢弃而强信号过饱和的潜在错误，而基于测序技术的ChIP-seq则能够对基因组区域中的组蛋白修饰进行精确的定量。

再者，就两种技术的覆盖范围而言，ChIP-chip受到芯片容量的限制，仅能测定预设的基因组区域，而ChIP-seq则可以覆盖绝大部分的基因组区域，因此在进行全基因组范围组蛋白

修饰测定时，ChIP-seq技术所需的样品量更少，且相对成本较低，这也是目前ChIP-seq逐渐代替ChIP-chip成为主要的组蛋白修饰测定技术的另一个主要原因。

尽管目前人们普遍使用二者测定基因组范围的组蛋白修饰，然而二者都存在潜在的某些缺点，如ChIP-chip可能发生探针与非特异性区域的杂交，而ChIP-seq的测序数据则受到GC含量的影响，新的更高效的更精确的组蛋白修饰测定技术对基因组范围的表观遗传修饰调控机制的研究是必要的。

三、基因组印记与人类疾病

基因组印记是指从父本或母本遗传得到的等位基因间存在表达上差异的一种现象。印记基因通常表现为单等位基因的转录沉默，另一个等位基因正常表达，并且具有组织特异性。基因组印记的失调会导致复杂的病理现象，研究发现印记基因与多种疾病的发生有关系，如癌症、生长和代谢失调、神经发育和认知行为失调等疾病。与疾病相关的印记基因的功能失调可以从印记起源的角度上被理解为是对进化压力的反映。印记基因的表达暗示了基因组内进化矛盾的结果，使得从双亲遗传获得的等位基因为了适应进化选择表现出等位基因的差异表达。

基因组印记的突变与一些复杂疾病有着密切的关系，如安琪儿综合征（Angelman syndrome），普瑞德威利综合征（Prader-Willi syndrome），贝-威二氏综合征（Beckwith-Wiedmann syndrome）。基因种系发育过程中印记擦除和获得的失调是引起这些综合征的主要原因。另外，研究显示*IGF2*缺失可导致上皮干细胞的数目增加而易患结肠癌。正因为显示单亲本孕体发育失败，很多印记基因在胎盘和胚胎中参与细胞分化和生长调控，也在神经过程和行为中起关键作用。因此，印记基因表达受到干扰将引起几种重要的生长行为综合征以及癌症。

目前人类和小鼠基因组的印记基因的准确数目和印记的范围还不是很明确。据估计它们的真实数目在100~2100左右。尽管这方面的研究在过去25年取得了巨大的进展，然而，起作用的印记的生物学功能图谱还没有完成，因此准确识别全部哺乳动物基因组印记基因是很必要的。

（一）印记基因的主要特征

印记基因分布在整个基因组。尽管有些是独立的有些成对存在，大部分在基因组上呈现成簇出现的现象，而且在人类和小鼠中保持结构保守。这些印记区域包含父本或母本都表达的基因（即非印记基因）在多数情况下至少包含一个非编码RNA。这些等位特异表达的印记基因（更广泛的可能是印记位点）受表观遗传修饰调控，其中DNA甲基化是主要因素。

研究发现，离散的协同印记的单等位表达的顺式作用区域即印记控制区（imprinting control region，ICR）全部是发生差异甲基化的区域（differentially methylated region，DMR），包括等位的遗传自父本和母本的DNA甲基化（被认为是种系的DMR）。在ICRs种系的DNA甲基化的获取需要DNA甲基转移酶3A（DNMT3A）和DNA甲基转移酶3 Like（DNMT3L）共同作用。一旦获得，DNA甲基化印记将在整个发育和成体所有体细胞谱系中维持。这些印记标记以

不同方式读出，确保适当的亲本等位特异表达。

（二）印记基因的识别

1. 基于基因表达分析发现印记基因　研究发现亲本等位基因不平衡表达适用于识别印记的位点。然而，设计这样表达需要满足几个重要的条件：①理想的cDNA资源需要区分父本和母本的等位基因；②cDNA资源应该代表蛋白质编码和非编码的基因（非编码RNAs）因为两者的表达都能被印记调控；③有的基因只在特定的组织和发育阶段印记；④确定这些基因需要区别它们与基因显示随机单等位表达。

第一次成功表达扫描是使用小鼠单亲本cDNA胚胎或胚胎成纤维细胞基于消减杂交和差异显示技术实施。随着高通量芯片技术的发展可以同时扫描几千个基因的表达。尤其是，一个大规模的芯片由全9.5dpc单亲本小鼠胚胎的27 663个full-length个小鼠cDNA通过比较基因表达水平用于识别印记基因。分析识别出多于2100个印记候选转录本（分别1403母本表达和698父本表达，包括56个非编码RNA）。

在一个最近的研究中，9.5dpc母本和控制胚胎的基因表达水平用Affymetrix GeneChip探针芯片比较，包含多于45 000个基因和ESTs。只有39个候选转录本（包括18个已识别的印记），识别为父本表达。然而，这些结果进行实验验证时，只有很少的候选转录本被证实印记而大多数显示非印记。

小鼠的种系带有特定的单亲二倍体（UPD）或者重复的印记染色质区域可以至少部分的克服这些缺陷。例如，Schulz等人利用UPD小鼠不同组织的cDNA芯片实验成功的识别了三个胎盘中母本表达的新的基因，新的4个大脑组织特异的父本表达转录本。

然而，基于UPD小鼠芯片扫描的局限性也是显而易见的。事实上，在所有的单亲本胚胎中，在这些胚胎中观察到印记的缺陷可能干扰非印记基因的表达，产生假阳性的印记基因。

另一种方法是使用有意义的SNP变体使得不但能建立给定基因的亲本起源的表达，也能扫描生理正常的'material'。这些可以通过小鼠不同株系的相互杂交实现。在人类中，为了破解复杂疾病的起源，国际HapMAp协会建立了单体型图谱数据库（http://hapmap.ncbi.nlm.nih.gov/）和千人基因组计划（http://browser.1000genomes.org/index.html）。已识别的人类的SNP在dsSNP数据库（http://www.ncbi.nlm.nih.gov/projects/SNP/），可以用于转录水平来确定特定基因或一组基因的等位基因的表达。几个研究用SNP特异芯片的方法研究人类组织和细胞系等位特异的基因表达。虽然不一定致力于新印记基因的预测，这些研究确定了几个已知印记基因的差异的等位表达和几个高置信度的新的印记转录本。

为了识别新的印记基因，Pollard 等人设计一个方法允许从随机单等位表达的基因中分别出真实的候选印记基因。在这项研究中，等位表达研究是通过来自67个不相关个体的SNP特异芯片。在这些基因中，显示差异等位表达的，真正的候选印记基因通过SNP相关等位相对于不同杂合子个体其他等位过度或不足等位表达来识别。因此，分析2625个人类基因确定了61个候选印记基因。其中15个实验验证印记，7个显示强的证据，但没有证实被印记。

由于此实验只覆盖了人类编码基因大约10%（不包含非编码基因），估计在淋巴细胞没有多于几百个印记基因。尽管接下来的芯片将提供更广泛的覆盖率，这种方法也有很大的局限性。分析是定制选择的，限定一定的基因组区和他们需要已知的SNP位置和转录序列。

此外，芯片分析不能提供一个已知基因在两个等位基因的可靠地定量的表达比率，因此不便于识别显示一个等位亲本表达偏差的印记基因。

最近计算机辅助深度测序方法的发展，二代测序技术，提供了有效的有希望替代芯片分析的方法。尤其是致力于识别新的印记基因的RNA测序（RNA-seq）方法。因为当应用于多肽cDNA资源时可以定量检测整个转录本的等位偏差。Wang 等通过对不同株系反交的新生小鼠大脑样本的转录本进行测序，识别的26个印记基因中有3个被确认是印记的。Babak等用一个简单的方法建立了小鼠胚胎9.5dpc的印记基因的图谱。

除了转录本分析，单等位基因表达评估可以通过研究等位特异的转录结合因子。在一个新颖的方法中，Maynard等调查人类肺成纤维细胞的RNA聚合酶Ⅱ的等位特异结合位点。他们设计ChIP-SNP方法，其中用SNP分析芯片来分析沉淀抗RNA聚合酶抗体的区域。通过这种方法，识别了已知的印记基因，包括microRNA簇邻近的*MEG3*等。

2. 基于差异甲基化的研究　甲基化胞嘧啶的全基因组图谱通过无偏的方法系统的识别DMRs获得。特别有趣的是最近发现的BS-seq方法能应用到复杂的人类和小鼠的基因组。进一步将此方法结合SNP数据应用于有效的识别新的候选的DMRs相关的印记位点。

另外，基因组范围DNA甲基化谱可以用于比较正常基因组和那些已知存在甲基化印记缺陷的基因组。这一方法最近成功的应用于分析一个患者血液样本多个印记缺陷和正常控制基因组CpG甲基化。接下来的亚硫酸盐处理，DNA杂交到用于分析多于14 000多个基因的CpG甲基化芯片。除了确定这些患者在已知印记DMRs的低甲基化，这一研究还可以确定新的候选DMRs。和这些区域相关的RB1显示在人类基因组中发生印记。

3. 基于染色质特征扫描——一种未来时代的方法　几种定位特异研究组蛋白修饰，其中进一步被全基因组分析支持，揭示了一种ICR特异的染色质信号的存在。尤其是，DNA甲基化的等位基因与被定义为抑染色质的组蛋白标记一致相关（如H3K9me3和H4K20me3）。相反的，非甲基化的等位基因被组蛋白修饰H3K4me2/me3标记，是典型的活性染色质。应用认了和小鼠细胞系的全基因组Chip-seq得到的组蛋白标记图谱，用机器学习的方法可以应用这些数据系统的识别这些特异染色质信号的区域特征。此外，ChIP-seq识别的染色质特征会以等位特异的模式被读出，用SNPs允许分配每个染色质修饰到特定的亲本等位，因此用于识别新的候选ICRs。

除了识别假定的ICRs，全基因组的染色质信号也可用于识别等位转录的区域。研究显示ChIP-seq结合等位差异的SNP信息，观察到H3K36me3等位不平衡，在一些印记位点和microRNAs标记与转录延伸相关。

通过获得不同表观特征谱交叉信息可以进一步改进这些方法。使用包括几个小鼠印记染色质区域的定制芯片，Dindot等人发现已知的ICRs是有特异的DNA甲基化谱与H3K9me3和H3K4me3重叠。通过这一方法他们识别了11个新的印迹控制区。通过使用一个简单的方法，Wen等人提出可以用H3K4me2，DNA甲基化和CTCF结合位点识别人类T细胞和永生淋巴细胞系的印记区域。

4. 通过DNA序列特征预测印记基因　已经识别的大量的印记基因为计算方法识别候选印记基因提供了条件。计算方法是基于识别的印记基因共有的特定的序列特征。尤其是，人类印记区域相对于非印记位点显著缺少SINEs。

第一次大规模基于DNA序列特征预测印记基因的是Luedi等人比较分析了44个已

知印记基因和530个状态没有经实验验证的基因(假定非印记基因)。分析包括几种重复元件家族的分布,转录因子结合位点和CpG岛。其中显著的特征,作者证实低密度的SINEs是印记区域一个显著的特征,并且相对于印记基因的方向有很高的区分价值。此外,内含子中的逆源性病毒和LINEs L1s也是预测印记状态的重要因子。随后,这些显著预测特征被用于训练分类器用于预测基因的印记和非印记状态以及印记的表达情况。

应用分类器分析全部23 788个注释的常染色体基因结果识别600(2.5%)个候选印记基因,384(64%)个预测为母本表达。相似的,将分类器用于人类,预测出20 770个基因中的156个,88个为母本表达。基因组中预测的小鼠和人类印记基因比率的差异可以通过更严格的方法来解释,但可能也是事实上人类基因组就是比小鼠有更少的印记基因。此外,两种物种的基因组预测的印记基因的数目都只限定在蛋白质编码基因。

最后,人类基因组中156个候选基因,只有*DLGAP2*和*KCNK9*,被实验证实为印记基因,并预测为父本等位基因表达。有趣的是,*KCNK9*,在之前对小鼠研究中显示大脑中母本印记特异表达。Luedi等预测的600个基因中的16个可能的候选印记基因被用E11.5小鼠胚胎实验检验。除了*KCNK9*,其他15个基因在这个发育阶段没有显示印记特异表达。

5. 基于表观遗传特征预测　扫描检测表观标记(如DNA甲基化和组蛋白修饰)在给定基因父本和母本的差异。也提供相应的策略识别新的印记位点。已知在ICRs有等位表观遗传差异,印记调控的关键区域。

ICRs是由DNA甲基化在父本或母本一方种系标记获得的(构成种系DMR)。除了DNA甲基化,ICRs在一些体细胞中也有差异的组蛋白修饰标记。扫描识别种系DMRs/ICRS的主要优势集中在他们可以在所有的细胞类型中执行,因为无论印记基因的的表达水平如何,它们的表观遗传标记覆盖在发育相关和成体细胞。另一方面,ICRs作为离散的元件通常控制成百上千个碱基最多十个基因的整个印记簇。

因此,这种扫描更倾向于识别几个印记基因的染色质区域而不是单个的印记基因。随之而来识别启动子区域的组织差异甲基化区域(T-DMRs)更容易显示单个的印记基因。

正如在转录组分析一样,过去几年经历了巨大的技术变革,有利于以一种无偏的方式显示全基因组的表观遗传特征。甲基化的DNA由抗—5mC抗体沉淀(MeDIP assay)或者甲基-CpG结合蛋白(MIRA assay)能进一步通过芯片杂交或深度测序方法分析。

另一个有意义的方法是重亚硫酸盐测序(BS-seq),其中甲基化依赖重亚硫酸盐保守DNA(甲基非甲基胞嘧啶差异)结合高通量测序全基因组单个碱基定量的图谱。相似的,全基因组的组蛋白修饰图谱通过特定的染色质免疫和嵌入探针芯片(ChIP-chip)或深度测序(ChIP-seq)获得。

四、常用的疾病表观遗传学数据库

随着高通量实验技术的不断推出及改进使表观基因组水平的数据与日俱增,研究人员测定了各种疾病状态下的基因组范围的表观遗传学修饰的图谱,如何存储如此众多且重要的数据并从中提取重要的信息成为表观遗传学研究的瓶颈,为了解决这些难题,研究结合生物信息学技术,构建了专门的疾病表观遗传数据库用于存储疾病相关的表观遗传学实验测

定的数据，并在数据库中开发了相应的功能分析模块以供科研人员进一步分析数据中的重要信息。

疾病表观遗传学数据库的构建和应用促进了表观遗传学的快速发展，有利于相关数据的重复利用。疾病表观遗传学数据库主要是用来存储疾病及其正常对照中的各种表观遗传学修饰(如DNA甲基化、组蛋白修饰等)的数据，例如，人类疾病甲基化数据库DiseaseMeth、人类DNA甲基化与癌症数据库MethyCancer、人类组蛋白修饰数据库HHMD。下面将简单介绍这两个典型的疾病表观遗传学数据库的网站及其使用。

(一) 人类疾病甲基化数据库(DiseaseMeth)

人类疾病和DNA甲基化的改变密切相关。人类疾病甲基化数据库(DiseaseMeth)是人类甲基化数据库中迄今为止收录各类实验测定的人类基因甲基化数据最为全面的数据库，该数据库旨在存在人类各种组织、细胞系在疾病等状态下的高通量和小规模实验的甲基化数据(图11-4)。

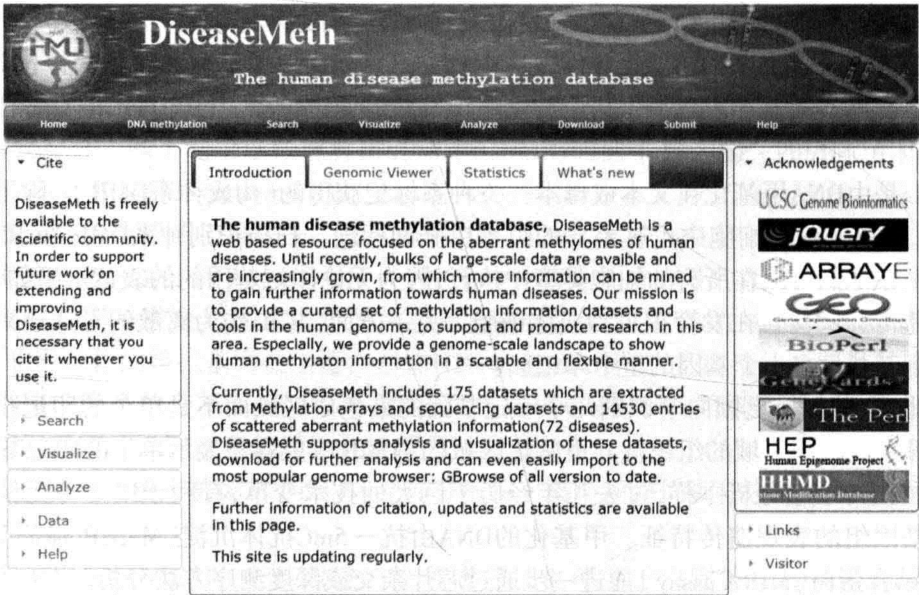


图11-4 人类疾病甲基化数据库(DiseaseMeth)首页

当前数据库收录72种疾病的超过14 000个条目的数据。该数据库提供在线查询、下载、分析和可视化等基本工具。可通过疾病类型、染色体位置、基因、细胞类型，实验技术等多种选项联合筛选，进而进行基因中心的甲基化分析。分析结果链接了可视化界面，提供了方便使用的基因组角度的视图。结果页面还链接到多个数据库，如HHMD、GeneCards、MethyCancer等。该数据库还内建了分析基因和疾病之间关系的分析工具，可方便地进行基因-基因、基因-疾病和疾病-疾病之间的相关性分析。同时，下载的数据可被其他软件如GBrowse识别，便利下游的功能研究。该数据库为甲基化研究提供了新的便捷工具，为阐释癌症的发生机制、筛选疾病相关的基因提供了便利的研究工具。更多DiseaseMeth详情可以访问其官方网站：http: //bioinfo.hrbmu.edu.cn/diseasemeth。

(二) 人类DNA甲基化与癌症数据库 (MethyCancer)

人类DNA甲基化与癌症数据库(MethyCancer)是第一个比较全面的人类癌症DNA甲基化数据库(图11-5),该数据库旨在研究DNA甲基化、基因表达与癌症间的相互作用,涵盖DNA甲基化、癌症相关基因、突变、癌症信息和CpG岛等信息,对这些不同数据类型之间的互联互通进行了分析和讨论。

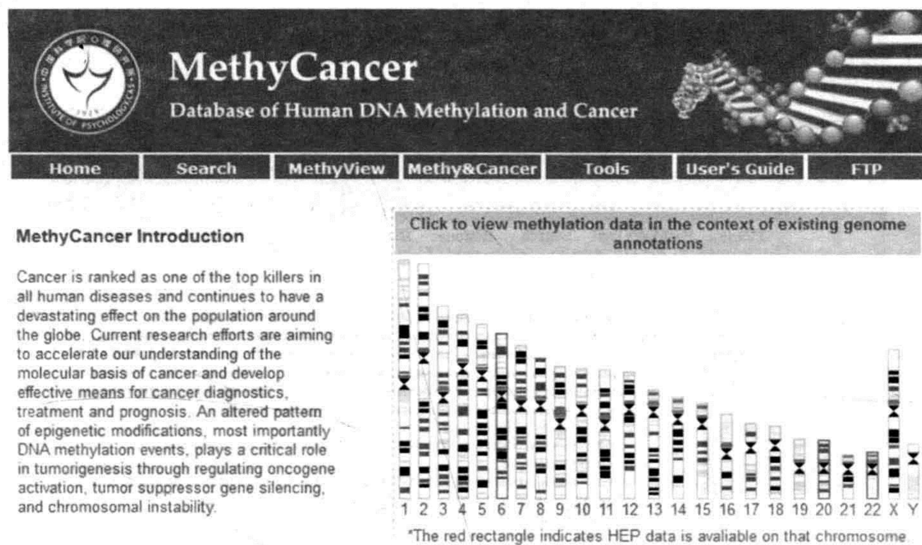


图11-5 人类DNA甲基化与癌症数据库(MethyCancer)首页

当前版本的MethyCancer存储了人类表观基因组计划(HEP)等测定的人类145个组织中的DNA甲基化数据,以及7075个癌症相关基因中的DNA甲基化模式,为进一步分析癌症中异常DNA甲基化提供了数据资源。MethyCancer的DNA甲基化搜索界面,可用来搜索用户感兴趣区域的甲基化模式。另外该数据库还提供了基因搜索界面、癌症搜索界面、序列搜索界面及重复序列搜索界面。MethyCancer还提供了搜索工具和可视化工具(MethyView)来帮助用户获取感兴趣的数据并在基因组的背景下查看DNA甲基化模式。该数据库加速了研究者对癌症发生的分子机制的阐明,并促进了癌症诊断、治疗机预后的有效手段的研究。更多MethyCancer详情可以访问其官方网站: http: //methycancer.psych.ac.cn/。

(三) 人类组蛋白修饰数据库

人类组蛋白修饰数据库(human histone modification database, HHMD)是人类组蛋白修饰数据库是迄今为止收录各种实验测定的人类基因组组蛋白修饰最为全面的数据库(图11-6),该数据库旨在存储人类各组织中的高通量组蛋白修饰数据,并提供各种癌症基因上的组蛋白修饰状态。当前版本的HHMD共涵盖了43种基于ChIP技术的实验技术测定的人类组蛋白修饰的大通量实验数据,并提供了通过文献得到的9种癌症相关的基因的组蛋白修饰的信息。

用户可以通过搜索相应的基因组区域中的组蛋白修饰,该数据库提供了五种搜索组蛋白修饰的方式,分别是组蛋白修饰类型、基因ID、功能注释、染色体定位、癌症类型。并可以

通过可视化组蛋白修饰的工具HisModView进行基因组水平可视化，在已有的基因组注释的背景下研究组蛋白修饰的分布、这些组蛋白修饰与DNA甲基化之间的关系，以及二者与相应基因功能元件的位置关系，据此来设计实验对感兴趣的区域进行湿实验研究。该数据库支持用户对搜索和可视化的结果进行下载，并且提供了处理基因组组蛋白修饰数据的Java程序。整个数据库体现了很好的交互性操作，对研究组蛋白修饰与其他表观遗传调控元件如DNA甲基化之间的相互作用关系提供了一个很好的平台，能够促进组蛋白修饰在染色质重塑、转录调控和人类疾病中作用机制的研究。更多HHMD详情可以访问其官方网站: http: //bioinfo.hrbmu.edu.cn/hhmd。

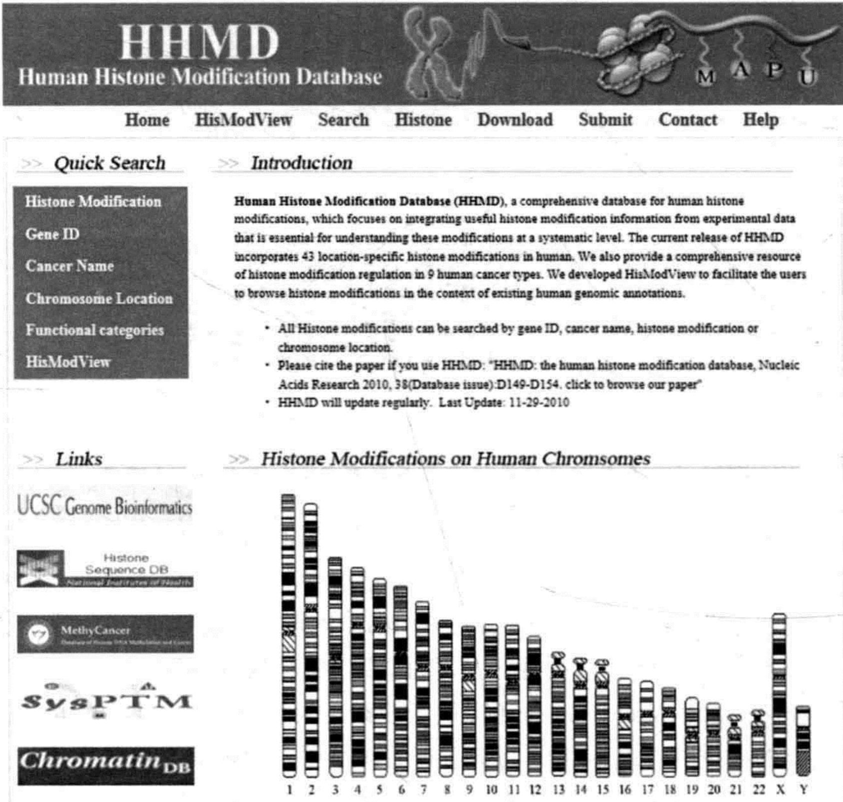


图11-6 人类组蛋白修饰数据库(HHMD)首页

第三节

表观遗传修饰谱分析

Section 3 Analysis of epigenetic modification map

表观遗传调控失调与许多疾病有着密切的关系。如表观遗传机制调控免疫系统的功能，当这一机制失调时，会引起类风湿疾病和系统性红斑狼疮的发生。同样，表观遗传调控大脑神经细胞的活性，精神性疾病如精神病和药物滥用反应与表观遗传改变相关。对于所有的常见疾病，肿瘤中表观遗传改变的作用研究得最详细。肿瘤抑制基因的表观遗传沉默是多种肿瘤的频繁发生事件，一些证据表明它是肿瘤发展的重要原因，如最近的研究显示DNA甲基化模式在癌细胞和正常细胞存在较大的差异，在正常细胞中，全基因组是高甲基化而在CpG岛中低甲基化。在癌细胞中则呈现全基因组的低甲基化和某些CpG岛的高甲基化，这则可能与DNA甲基化调控致癌基因和抑癌基因的相对表达水平相关。癌细胞和成体胚胎干细胞之间有着相似的表观遗传特征，暗示着表观遗传失调使细胞行为会向肿瘤细胞方向发展。

一、基因组范围内疾病差异甲基化区域筛选

(一) 差异甲基化区域的生物学意义

DNA甲基化的差异在发育过程和疾病的发生发展过程中扮演着重要的角色。基因的差异甲基化区域（DMRs）是指甲基化模式发生改变的区域，研究发现它们可能是调控基因转录的功能区域。大量的研究发现在人类的各组织间存在组织差异甲基化区域（T-DMRs），癌症中存在癌症差异甲基化区域（C-DMRs），对差异甲基化区域的识别可促进人类基因组中表观遗传变异的生物学意义的研究。

目前普遍认为DNA甲基化参与细胞增殖和分化，不仅在发育阶段发现了发育差异甲基化区域（D-DMRs），而且在重编程的过程中也发现了重编程差异甲基化区域（R-DMRs），这些R-DMRs与T-DMRs和C-DMRs也有着高度的重叠。

此外，随着年龄的增长表现出个体内的差异甲基化区域（Intra-DMRs），以及在多个个体间的差异甲基化区域（Inter-DMRs）。通过对DMRs的研究能够更深入地了解DNA甲基化与其他表观遗传调控因子协同调控基因功能的具体机制。

(二) 差异甲基化区域的筛选方法

目前关于差异甲基化区域相关的研究中，已有了一些计算方法用来从实验数据中识别

差异甲基化区域。在最初关于差异甲基化区域的研究中，由于实验技术的限制，人们只关注少数几个基因在样本间的DNA甲基化差异，如Shen等通过限制性内切酶方法测定一簇β-珠蛋白基因中的DNA甲基化，然后通过观察DNA甲基化在各组织中的差异来分析组织特异的DNA甲基化。

随着高通量实验技术的不断进步，在过去的几年里，一直在努力开发计算的方法来识别DMRs。为了在两个样本间筛选DMRs，Bibikova等人对人类胚胎干细胞和正常完全分化细胞中的甲基化水平进行*T*检验。在另一项研究中，为了进行配对组织比较，使用平均甲基化的差异以及*z*值来衡量差异甲基化。然而，这两种方法并不适用于多于样本的情形。

在处理多样本的数据时，有两种筛选DMRs的统计学方法：Byun等使用的方差分析（ANOVA）和Eckhardt等使用的Kruskall-Wallis检验。Byun等使用的前提是数据服从正态分布，但是这个假设在服从双峰分布的甲基化数据中是不存在的，所以该方法对于分析DNA甲基化数据有些受限。

Eckhardt等使用的Kruskall-Wallis检验不依赖于数据分布，比前者更适合用来筛选DMRs，但是这种统计方法利用的数据中甲基化状态的排秩，对于排秩顺序相同的甲基化区域，能够给出相同的结果，但是并没有考虑到甲基化波动范围带来的影响，如波动范围大的区域应该有更大的差异，因此这种方法可能丢失原始数据中丰富的数字信息。

除了统计学方法，还有两种非统计学的方法，Fan等将在所有组织中的甲基化程度都大于50%的区域定义为甲基化区域，都小于50%的区域定义为非甲基化区域，其他区域则为DMRs，显然这种方法忽视了甲基化程度在50%附近的区域的正确分类。

作为这种方法的改进，Rakyan等提出将某个组织中超高甲基化T-DMRs定义为在该组织中甲基化程度大于60%且至少在三个其他组织中的甲基化程度小于40%的区域，超低甲基化T-DMRs则为该组织中小于40%且在至少三个其他组织中大于60%的区域，剩余的区域为非差异甲基化区域。

有人利用改进后的香农信息墒开发了一个新的方法QDMR，并提供了界面友好的简单易用的JAVA软件，对各种不同样本间的甲基化差异进行定量化，并从基因组范围内筛选出差异甲基化的基因组区域。基于之前用于筛选差异表达基因的信息熵的方法，通过输入值校正和熵值校正两步改进，用来筛选差异甲基化区域，这是信息熵理论第一次用来筛选某种因素在不用样本间的差异程度，并且是第一次用来筛选差异甲基化区域。基于信息熵的方法不仅能够衡量每个基因组区域在不同生命状态间的定量差异程度，而且能够根据这种差异程度来对所有区域进行排秩或者分类，除此之外，还可以指出差异甲基化区域在哪个生命状态下特异的发生高/低甲基化，从而可以研究不同类别的区域在不同生命过程中所扮演的不同角色。

【例11-1】QDMR方法筛选差异甲基化区域

表11-4给出了40 437个人类基因组区域在16个组织中的甲基化状态，利用QDMR软件对组织间的甲基化差异进行定量，并筛选差异甲基化区域DMR。

表11-4　人类16个组织中的甲基化数据

区域	组织类型															
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16
ROI000001	20	31	31	34	19	24	20	18	34	31	29	13	27	34	26	17
ROI000002	60	56	63	63	57	61	58	59	56	55	54	63	56	58	59	61
ROI000003	64	55	63	61	61	62	58	60	54	52	54	69	60	58	61	62
ROI000004	35	40	43	51	38	35	43	38	46	41	53	23	51	53	37	48
ROI000005	40	36	42	44	37	46	42	38	39	41	43	28	45	40	36	45
ROI000006	34	27	34	29	28	25	28	24	36	32	41	18	35	29	27	30
ROI000007	27	29	36	30	25	31	26	18	30	29	42	12	37	32	29	29
ROI000008	71	57	74	67	72	78	75	76	61	65	67	67	71	62	71	59
ROI000009	66	48	74	61	44	66	55	51	57	63	67	31	67	59	76	55
ROI000010	24	22	27	26	21	23	22	21	24	27	29	20	31	25	24	23
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
ROI040437	50	44	60	58	55	54	51	42	51	62	65	34	54	52	29	55

注：数据来源：Rakyan VK.An integrated resource for genome-wide identification and analysis of methylated regions (tDMRs).Genome Res.2008；18,1518-1529。

1. 甲基化差异定量

$$H_Q=-\sum_{s=1}^{N}p_{s/r}\log_2(p_{s/r})\times\left|\log_2\left(\frac{\max(m_{r,s})-\min(m_{r,s})}{MAX-MIN}+\varepsilon\right)\right| \qquad (11\text{-}1)$$

其中N为样本总数，$m_{r,s}$为区域r在样本s中的甲基化值，$\max(m_{r,s})$、$\min(m_{r,s})$、MAX和MIN分别为区域r各样本中的最大甲基化值、区域r在各样本中的最小甲基化值、整体最大甲基化值和整体最小甲基化值，ε为非常小的数，T_{br}为一步双加权的距离(用于避免原始熵公式对特异高甲基化的偏好性)。

$$p_{s/r}=\frac{\left|m_{r,s}-T_{br}\right|}{\sum_{s=1}^{N}\left|m_{r,s}-T_{br}\right|} \qquad (11\text{-}2)$$

本例中以区域ROI000001的熵值H_Q为8.457。对数据中的每个区域都利用上述公式进行甲基化差异定量(图11-7)。

2. 差异甲基化区域筛选

为了基于定量的甲基化熵筛选组织间差异甲基化的区域，该方法又基于概率论是思想，根据假设$\frac{m_{r,s}-Mean}{MAX-MIN}$（中$Mean=\frac{1}{2}(MAX-MIN)$）服从均值为0标准差为SD的正态分布，确定相应的SD值即可模拟随机的甲基化谱，将随机甲基化谱获得的熵值的平均值作为筛选差异甲基化区域的阈值，该方法默认取SD为0.07(意味着同一个区域95%的样本中甲基化值

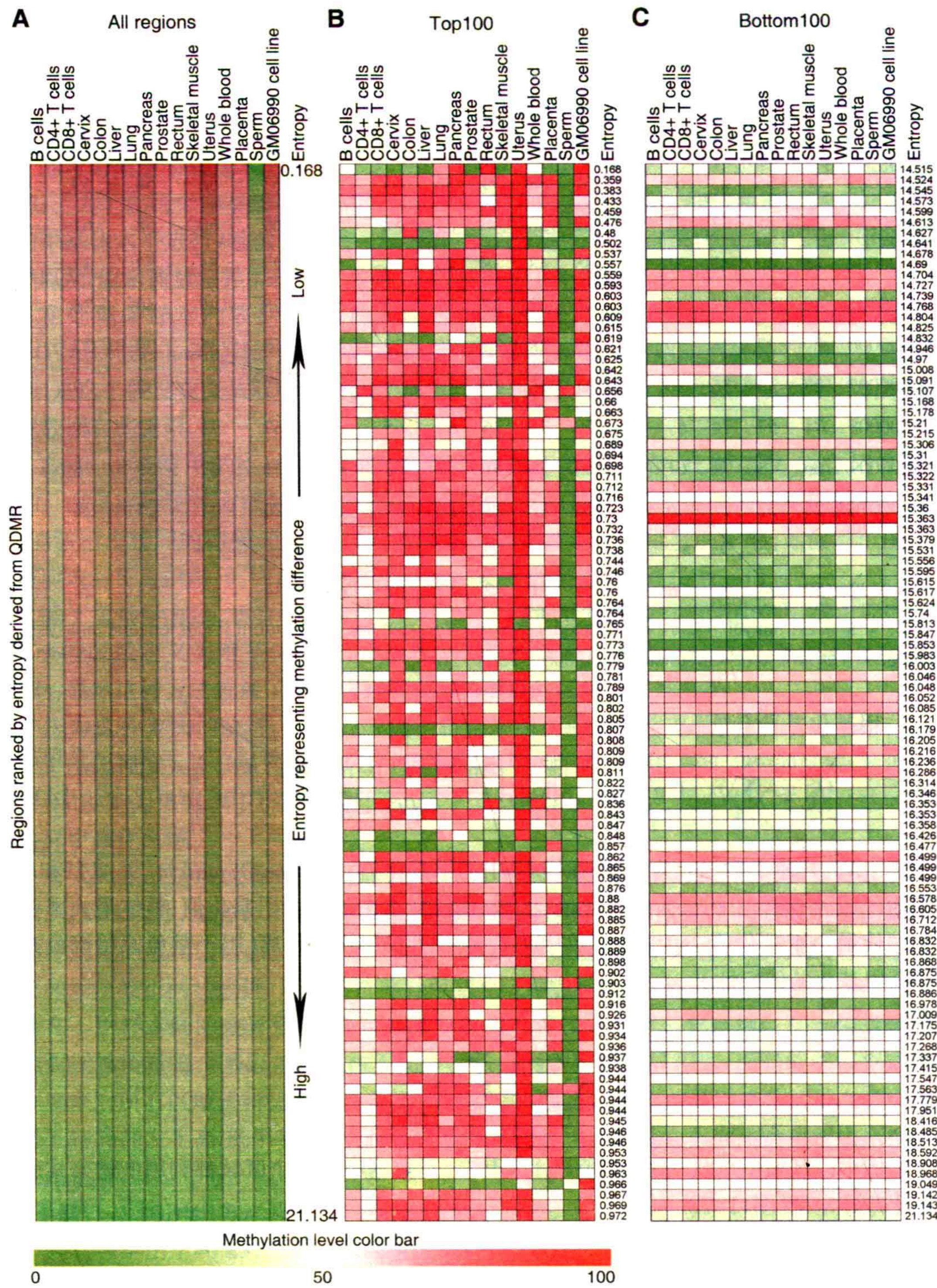


图11-7 人类16组织间甲基化差异定量

A.为40437个区域的甲基化热图，熵值由小到大排序。B.为A中上方熵值最小的100个区域的热图，C.为A中下方熵值最大的100个区域热图

来 源 于：Zhang Y. QDMR: a quantitative method for identification of differentially methylated regions by encropy.Nucleic Acids Res.2011; 39 : 58.

在36到64之间)，获得筛选组织间差异甲基化区域的阈值5.326。根据该阈值从40 437个区域中筛选出10 651个组织间差异甲基化区域(图11-8)。

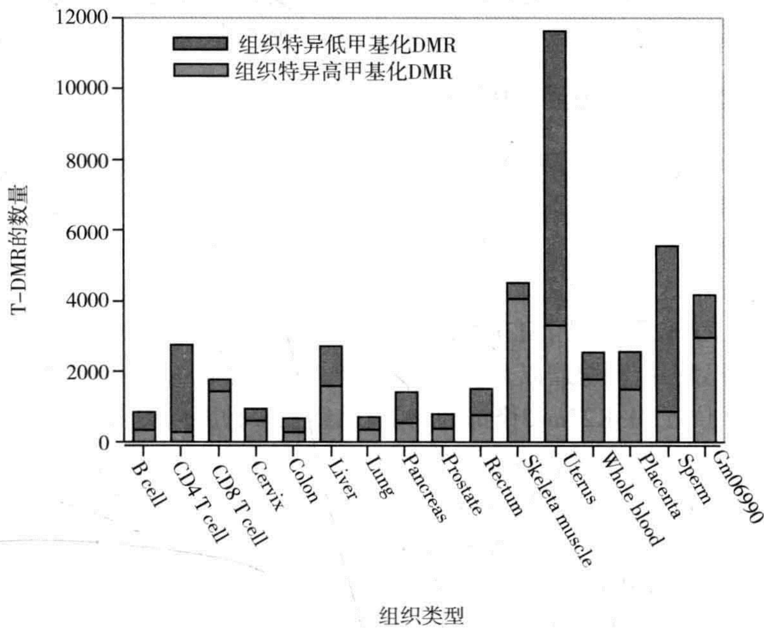


图11-8 人类16组织间差异甲基化区域及组织特异性
来 源 于: Zhang Y. QDMR: a quantitative method for identification of differentially methylated regions by encropy.Nucleic Acids Res.2011; 39 : 58.

3. 样本特异性衡量

对于基于以上阈值筛选的差异甲基化区域，该方法利用差熵定义了各样本中绝对特异性测度:

$$CS_{r/s}=\begin{cases}\Delta H_{r/s}\times sign_{r,s}, & \Delta H_{r/s}>0\\ 0, & \Delta H_{r/s}\leqslant 0\end{cases} \qquad (11\text{-}3)$$

其中$\Delta H_{r/s}=H_{Q/\bar{s}}-H_Q$，其中$H_{Q/\bar{s}}$为不含样本s时计算的熵值。用来衡量差异甲基化区域在各样本中的甲基化特异性。大于0的CS表示该DMR在该样本中特异高甲基化，小于0表示特异低甲基化。利用方法衡量以上筛选出的10 651个T-DMR的组织特异性(图11-8)。

二、组蛋白修饰的改变与人类疾病

(一) 组蛋白修饰改变参与疾病发生

组蛋白修饰在基因表达的调控中具有重要的作用，疾病状态下异常基因表达可能是导致疾病发生的原因之一。最近的研究已经表明癌症中组蛋白修饰H4K16ac和H4K20me3的全局性的缺失，而之前的研究已经表明这些修饰出现在整个基因组中，特别是覆盖在重复序列中的DNA低甲基化区域。相反，还有研究在特定基因启动子区发现了H3K9ac、H3K4me2和H3K4me3的缺失，以及H3K9me2、H3K9me3和H3K27me3的获得，这些组蛋白修饰的改变

能够沉默关键肿瘤抑制基因从而促进肿瘤发生。

另外发现癌症中DNA甲基化标记的基因和多梳家族蛋白（polycomb group protein，PcG）基因高度重叠，之前有报道称多梳家族蛋白经常与抑制性组蛋白修饰H3K27me3共定位，因此可以推测某些基因在癌症中被特定的抑制性组蛋白修饰靶向，导致其沉默。

（二）组蛋白修饰酶介导组蛋白修饰改变

在组蛋白修饰发生改变的过程中，组蛋白修饰相关的酶起了关键作用。通过催化酶的平衡对组蛋白修饰进行双向调控，这是翻译后修饰的典型特征。在疾病状态下催化酶的失衡将导致表观遗传修饰的异常，从而导致疾病相关基因的表达和功能异常。

最近的研究显示这种失衡在人类和小鼠的肿瘤生成具有重要的作用。例如，组蛋白甲基转移酶（HMTs）EZH2能够催化H3K27me3，该酶的过表达能够促进肿瘤生长，这种肿瘤包括如黑色素瘤、淋巴瘤、前列腺瘤以及乳腺癌。同时，癌症中H3K27me3的组蛋白去甲基化酶（HDMTs）的失活则导致H3K27me3修饰的增加，已经在多种肿瘤中发现了该现象，如多发性骨髓瘤，食管鳞状细胞癌和肾细胞癌等。

此外，研究还发现H3K9me3甲基化转移酶SUV39H参与肿瘤的发生和发展，该酶在小鼠中的缺失导致染色质不稳定并促进肿瘤的生成。除癌症中组蛋白甲基化转移酶外，组蛋白乙酰化转移酶（HAT）和组蛋白去乙酰化酶（HDAC）也导致大量的基因特异组蛋白乙酰化改变。研究发现在急性白血病中HDACs和HMTs的异常能够介导基因表达的沉默。对各种组蛋白修饰酶的平衡机制的研究将有助于解释组蛋白修饰参与疾病发生发展的作用。

三、人类疾病相关的基因组印记分析

（一）基因组印记分析的意义

通常情况下，印记基因的印记发生过程通过逐渐的亲本固定的随机单等位进化和表达来实现，如果两个等位基因都沉默则产生致死效应，固定的印记基因单等位表达确保了精确的蛋白质水平调节胚胎和产后生长以及行为、代谢。很多研究已经揭示遗传和表观遗传的机制调节印记基因的表达，发育中印记基因显示部分的基因簇共调控，印记基因位点也显示彼此的位置影响一些表观遗传状态的调控。这些发现表明印记基因在调控生长代谢和行为上可能是功能协调互作的。基因调控水平的协调又能控制基因编码蛋白质之间蛋白质-蛋白质的相互作用。因此系统的整合大规模的蛋白互作和基因表达数据可以从一个新的视角发现印记基因功能，这对正常发育和一些复杂疾病失调是很有意义的。

（二）利用大规模蛋白质互作网络分析印记基因

Sandhu等人应用公共的PPI网络，疾病和功能数据系统水平分析印记基因的功能，揭示了印记基因“products”和他们的互作“partners”有广泛的相互作用，共同构成一个生长与代谢相关的子网，在维持人类互作组拓扑和功能稳定性起重要作用。

【例11-2】基于网络识别印记基因的过程

（1）构建和分析IGPN（Imprinted Gene-products and Partners Network）。把已知的人类207个实验发现和预测的印记基因（来自于geneimprint resource，http://geneimprint.com）映射

到人类整合的蛋白质互作网络。通过检测IGPN和随机产生网络的全局和局部聚类系数确认印记基因产物和搭档的共同性。通过分析证实了IGPN代表人类互作组中紧密结合的网络,起主要调节功能的印记基因参与其中。

（2）评估IGPN到HIN的无偏好性和全局贡献。计算全局中心分数命名为IGPN和HIN的"betweenness"和"closeness"。Betweenness中心是根据网络中通过一个特定顶点的左右最短路径的数目,而closeness代表网络中一个顶点到达任何其他顶点的步数,这说明顶点的全局相关性。分析显示IGPN顶点与HIN和随机网络相比有更显著多的中心。由于遗传的突变可能导致完全的或部分的蛋白质互作缺失。基因表达失调在另一种程度上可能影响互作的partners的绑定和非绑定程度。非随机更高的顶点代表可能经验与疾病相关的在基因水平的错误或者扰动,可能意味着更多的IGPN功能的易受攻击性。

（3）探讨OMIM数据库中的疾病条目。发现IGPN映射的253个基因与疾病相关。疾病相关的基因与其他IGPN相比有更高的中心。此外,顶点中心和定点疾病数目正相关,强调这种更高的拓扑中心与功能本质的相关性。

（4）疾病与印记基因的关系。通过基因集富集分析（gene set enrichment analysis）,进行几次随机扰动确定了特定表达表型的非随机相关基因集,显示IGPN基因子集在疾病表型中显著失调。尤其是IGPN基因子集在ALL/AML,精神病和糖尿病中上调,在自闭症中下调。在哮喘和老年性痴呆中这种扰动是很微弱的。共88%的IGPN基因和70%的印记基因在疾病表型中是失调的,因此发现复杂疾病已知与印记基因相关。

进一步分析发现关键的信号和转录调节因子如T53, NFKB2, HDAC1, SMAD1-4, EGFR, TGFB1, EP300是高度的中心节点并与多数疾病的一般扰动相关。IGPN蛋白质更高的中心和连通性直接说明他们的生物学意义。这将可能包括全全局的调控子、适应子、效应子和调节分子。例如, GRB2, TRAF6/2是信号适应子, SMAD2/3, EGFR是重要的信号转导因子, TP53是主要的细胞周期调控因子, IKBKE激酶和EP300是转录因子结合在不同的增强子上,最主要的IGPN中心hubs在HIN分别有398,358,323,264,210 和196个互作蛋白质。IGPN中的印记基因GNAS（Ga 信号调节因子）,TP73（细胞周期,印记状态冲突）,INS（信号）,UBE3A（趴3泛素调节退化因子）, DCN（肿瘤增长抑制因子）, GRB10（信号转导）和 NDN（神经发育通路调节）最高互作节点分别有53,30,29,27,26,25和20互作partners。因此,印记基因本身并不是突出的hubs,尽管他们连接到HIN非随机的中心hubs和与IGPN在一起。印记基因可以作为VIPclub的类似物,因为他们自己不与很多的partners相连,但是多是影响partners。

四、常见的疾病表观遗传修饰数据分析软件

高通量技术产生的表观遗传数据直接促进了各种表观遗传软件的开发,这些软件也促进了表观基因组学研究的不断深入。为了从基因组水平研究表观遗传学修饰,需要开发对表观遗传修饰进行功能基因组分析的软件。目前,可用的软件中包括用于基因组筛选差异甲基化区域的软件（QDMR）,用于基因组CpG岛预测的软件（CpG_MI）以及（表观）基因组分析软件（EpiGraph）等。下面将对这三个计算表观遗传学软件的应用进行简单的介绍。

(一) 基于信息熵定量筛选差异甲基化区域软件(QDMR)

基于信息熵定量筛选差异甲基化区域软件(QDMR)是一个界面友好的定量筛选基因组多样本间差异甲基化区域的软件(图11-9)。随着基因组范围内DNA甲基化测定技术的不断进步,目前产生了大量的不同细胞/组织中的DNA甲基化数据,QDMR基于信息熵理论可以定量的筛选这些不同细胞/组织间的差异甲基化区域。

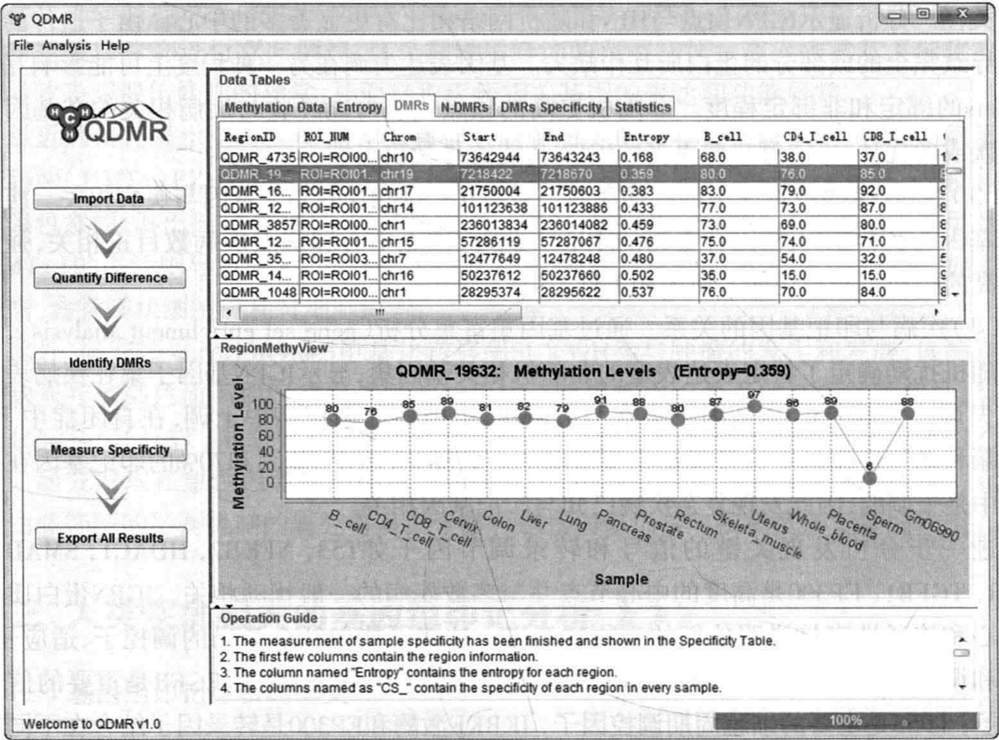


图11-9 基于信息熵定量筛选差异甲基化区域软件(QDMR)

该软件首先利用信息熵对各细胞/组织间的甲基化差异进行定量,然后再通过恰当的阈值筛选差异甲基化区域,并计算每个差异甲基化区域在各细胞/组织中的甲基化特异性。同时,该软件不仅对处理的数据提供了可视化,还提供了UCSC基因组浏览器的连接,便于用于查看差异甲基化区域附近的基因组信息及其他调控元件信息等。该软件适用于所有能够转换为0到1的连续数表示甲基化状态的实验技术。基于Web的和本地化的QDMR软件的下载及更多说明可以访问其官方网站http: //bioinfo.hrbmu.edu.cn/qdmr/。

(二) 界面友好的(表观)基因组分析和预测软件(EpiGraph)

EpiGraph,是一个界面友好的(表观)基因组分析和预测的在线软件(图11-10)。EpiGraph可用于复杂的基因组和表观基因组数据集的生物信息学分析,重于以组为单位的涉及两分类问题的基因组区域的分析。使得生物学家可以在脊椎动物基因组和表观基因组数据集中发现隐含的关联。

EpiGraph根据用户提交的一组基因组区域,测试多种属性(包括DNA序列,染色质结构,表观遗传学修饰以及进化保守)是否在这些区域中富集或缺失。此外,EpiGraph将会以预测

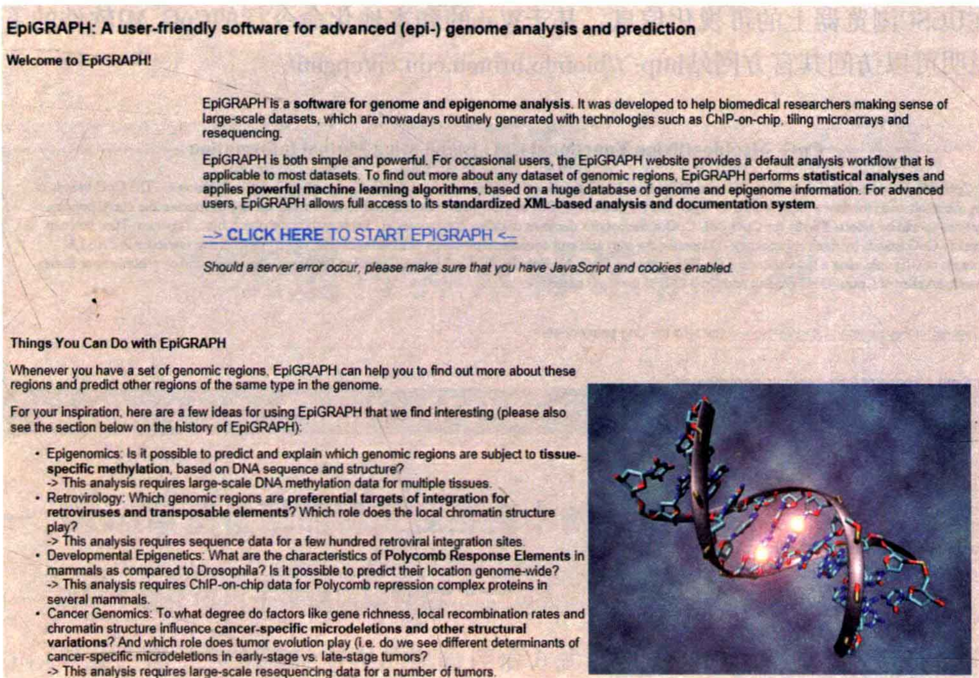


图11-10　EpiGraph在线软件

的方式鉴别相似的基因组区域。EpiGraph为用户提供了统计分析的在线服务，用户只需要简单地点击就可以完成对结果的分析，如网站相关的文献提供的单等位表达的例子中用于两组数据比较的箱式图。EpiGraph解决了两个基因组生物学的普遍任务：一个是发现一组特定生物学作用的基因组区域（例如实验定位的增强子，表观遗传调控的热点或表现出疾病特异异常的位点）和从公共数据库中得到的大量的基因组注释数据的新的关联。另外是评价是否可能鉴别具有相似作用的额外的区域，而不必进行进一步的湿实验。EpiGraph在线软件服务及更多说明可以访问其官方网站：http: //epigraph.mpi-inf.mpg.de/WebGRAPH/。

（三）基于互信息识别基因组功能CpG岛的软件（CpG_MI）

CpG_MI是基于互信息识别基因组功能CpG岛的软件（图11-11）。该方法不依赖于传统方法对CpG岛长度的限制，与之前用来识别CpG岛的方法相比，有着更高的精度，且识别出来的CpG岛大部分与组蛋白修饰区域相关。由于该算法只依赖于基因组CpG二核苷酸的分布，分析得到其他的脊椎动物基因组的CpG二核苷酸均服从相同的指数分布，因此可以将此算法推广到其他基因组中CpG岛的预测。

CpG_MI提供了在线的和本地化的两种方式为用户提供CpG岛预测服务。在线服务主要是识别单个基因组区域中的CpG岛，提供了两种提交数据的方式：一种是提交基因组一段序列的位置信息，算法将在后台访问UCSC数据提取相应物种基因组区域的序列；一种是直接输入已经获取的FASTA格式的序列或上传FASTA格式的文件。如输入人类1号染色体10 014 500到10 036 800的基因组位置，结果页将列出这段区域中的每个CpG岛的位置、长度、CpG数目、GC含量和序列信息，将之下载到本地进行进一步的分析。此外，本地化软件能够对各哺乳动物全基因组的CpG岛进行预测。CpG_MI提供了10种哺乳动物基因组中预测的CpG岛的下载

以及在UCSC浏览器上的可视化信息。基于Web的和本地化命令行的CpG_MI软件的下载及更多说明可以访问其官方网站http: //bioinfo.hrbmu.edu.cn/cpgmi/。

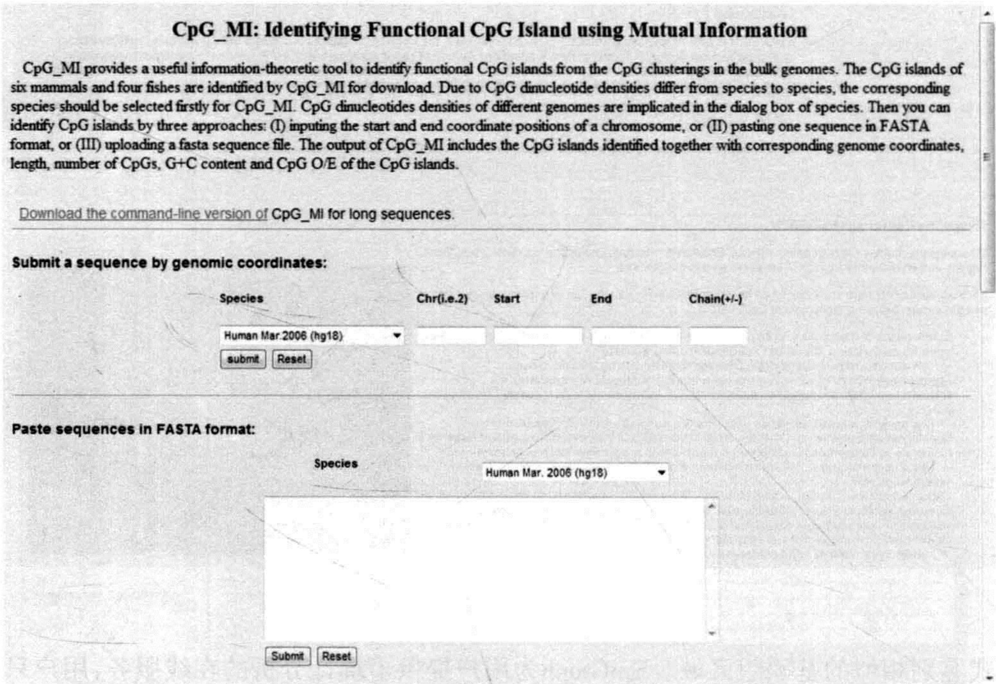


图11-11 CpG_MI在线软件

第四节

表观遗传异常标记物识别

Section 4 Identification of aberrant epigenetic biomarkers

一、DNA甲基化谱的特征在疾病中的应用

(一) DNA甲基化应用于疾病风险评估

DNA甲基化通常存在两种潜在的风险评估方式即组成性的异常DNA甲基化检测和能够预兆疾病发展的获得性变异的检测。前者与表观变异的隔代传递相关。尽管表观遗传标记的复位发生在种系中，使得在亲本和后代之间的表观遗传的遗传率高度变化，但是组成性的表观遗传变异在某些个体中是明显的，能够遗传的或者是一种后天种系缺陷。解释这种现象的临床实例是HNPCC（the autosomal dominant hereditary nonpolysis colorectal syndrome）综合征，受累个体在较早的年龄就高度易患结肠癌和子宫内膜癌。潜在的基因包括*MLH1*，*MSH2*，*MSH6*和*PMS2*。值得注意的是一小部分患有HNPCC的个体中发现在这些基因中没有一个发生序列突变，而发现*MLH1*或*MSH2*启动子甲基化也出现在正常组织中，包括循环血白细胞。在*MSH2*的研究中，发现*MSH2*邻近基因*TACSTD1*的突变，通过启动子和相关的DNA超甲基化导致异常转录。*MLH1*中没有发现类似的突变，因此呈现出一种罕见的遗传种系缺陷。

组成性表观遗传变异（epimutations）也能够偶然发生在邻近启动子区域并导致获得DNA甲基化倾向的以单碱基多态性形式的遗传变异产生。这种现象倾向于通过破坏反式作用保护蛋白的结合而发生，如Sp1。因此，表观遗传修饰的隔代遗传力能够通过顺式调控之间或者表观遗传自身的传送而形成，但是家族性癌症倾向很少仅仅有表观遗传现象相关的情况。

癌症风险评估的第二个方法是基于正常和倾向性组织的甲基化研究来检测获得性表观遗传变异。例如，在肺癌中，吸烟者的瘤前病变发现*p16*的甲基化，而在非吸烟者中没有发现该基因的甲基化。因此，*p16* 甲基化与其他基因结合（如*p14*，*p15*，*E-cadberin*和*RASSF1A*）能够被用作生物标记评估患者患有肺癌的风险，并且可通过检测唾液中甲基化实现。在一项研究中通过对98个疾病样本和92个对照样本的唾液中DNA甲基化的分析，发现14个基因的启动子甲基化可用于肺癌风险的评估，6个基因的启动子高甲基化被发现与超过50%的肺癌发展风险相关，这6个基因中的3个或更多基因的共同高甲基化使肺癌发生几率达6.5倍，灵敏度和特异性在65%的范围之内。值得注意的是在肺癌的临床发生之前的几年间，在唾

液中可探测到p16和MGMT高甲基化。

另一个例子，在结肠癌患者间发现癌组织和癌旁组织中同时出现*IGF Ⅱ*（insulin-like growth factor Ⅱ）这个基因的印记缺失。*IGF Ⅱ*的印记缺失在外周血淋巴细胞中也被发现，通过测定*IGF Ⅱ*的印记状态能够预测结肠癌发生的风险。在结肠中，正常组织中年龄相关的甲基化能够被打上标记癌症风险相关的区域缺陷，并且这个区域能够成为一个有用的生物标记。由于DNA甲基化能够通过药物介入而发生去甲基化，因此在肿瘤发生前的阶段的检测能够为癌症的预防策略打开一扇门，通过被动地对被测试组织严密的检测（连续的结肠镜检查/支气管镜检查法、成像研究等），或者主动地使用低甲基化药物或染色质重塑因子尝试恢复恶化前的表型。

（二）DNA甲基化的改变作为诊断的标记

在临床环境中的活检样本或体液中检测出异常DNA甲基化能够作为诊断生物标记，如血清、唾液、洗胃液、痰液、尿液或粪便。例如，*GSTP1*在人类前列腺癌组织样本中发现启动子区域的高甲基化，并且在对86个样本前列腺癌的研究中，检测出活组织样本中恶性肿瘤出现的敏感度为92%和特异性为86%。相似的，在粪便样本中波形蛋白甲基化的出现被发现在诊断结肠癌方面敏感性为46%和特异性为90%。

单一标记特异性的缺陷能够通过使用一组异常甲基化基因来补偿。来自175个患者和94个对照的尿沉积物DNA的检测中发现有9个基因甲基化能够预测膀胱癌，敏感性为82%，特异性为96%。使用DNA甲基化作为疾病诊断和评估的生物标记的一个局限性是来源于癌前病变的异常甲基化或者年龄相关的异常甲基化的可能性。目前公布的研究表明这是一种较好的阳性预测值的方法，有很高的敏感性和特异性。

（三）DNA甲基化谱的特征在疾病分型中的应用

表观遗传调控在癌症的发生和发展过程中所起到的作用已经广泛被关注。CpG的甲基化是在哺乳动物基因组中是最典型的表观遗传变化。通过CpG岛超甲基化导致的基因沉默是肿瘤中常发生的事件。此外，特殊基因的超甲基化，例如，在结肠癌的研究中，*ERα*、*MYOD1*和*N33*常常发生在老龄化个体的结肠组织中。因此，表观遗传变异的早期发现提出这样一个假说，它们允许随后影响肿瘤发生和发展的遗传和表观遗传变异的积累。重要的是，某些个体表现出有倾向性的异常启动子超甲基化，包括一些肿瘤抑制基因。这种现象，被称为CpG岛甲基化显性CIMP，提供一个导致结肠癌的另外的途径。一些研究已经将CIMP与遗传特征和临床特征相关联，包括*BRAF*和*KRAS*高的突变率，*p53*低的突变率，具体的组织学，家族性事件和不寻常的临床事件。

Shen等人分析97个原发的结肠癌患者的遗传（*BRAF*，*KRAS*，*p53*和微卫星不稳定性）和表观遗传变异（27个CpG岛启动子区的DNA甲基化）。基于表观遗传的DNA甲基化谱和遗传的基因表达谱的两个聚类分析识别出带有截然不同的分子特征的亚型。DNA甲基化的无监督分层聚类识别出三个结肠癌的分类，分别为CIMP1、CIMP2和CIMP-negation。在遗传学上，这三个分类与三个截然不同的基因表达谱相一致。CIMP1以*MSI*（80%）和*BRAF*（53%）的变异和少量的*KRAS*（16%）和*p53*（11%）变异为特征。CIMP2与92%的*KRAS*变异和少量的*MSI*、*BRAF*、*p53*变异（0，4%，31%）有关。CIMP-negative有高的*p53*变异的比率（71%）和较

低的*MSI*、*BRAF*、*KRAS*变异的比率（12%，2%，33%）。基于表观遗传和遗传特征的聚类也识别了三个分类，这三个分类与之前的分类有很大程度上的重叠。这三个分类不依赖于年龄、性别和阶段，但CIMP1和CIMP2在近端肿瘤中更常见。

（四）DNA甲基化改变与治疗反应的评估

甲基化模式能够用于评估临床效果和对化疗治疗药物的应答。通常，高DNA甲基化模式与较差的预后相关，如在肺癌或骨髓增生异常综合征。一项51个非小细胞肺癌I阶段样本和116个对照样本的实验中，在组织和淋巴结样本中检测7个基因的启动子甲基化状态与NSCLC复发的相关性。这些基因中的4个基因的甲基化（*P16*，*CDH13*，*RASSF1A*和*APC*）表明与肿瘤发生的独立相关性，*P16*和*CDH13*的甲基化被发现在训练集和结合训练集和验证集的集合中分别是15.5和25.25的比值。相似的，通过一组对10个基因的研究评估，当比较于低水平甲基化的患者，高水平甲基化的MDS患者被发现有较短的中值存活数（12.3个月 vs 17.5个月，p=0.04）和较短的无进展生存期（6.4个月 vs 14.9个月，p=0.009）。

然而，在一些例子中，强烈的高甲基化定义了一个特别的癌症子集，有一个较好的预后。结肠癌的研究中，多基因的CpG岛甲基化被命名为CIMP，与*MLH1*甲基化相关，导致了较好的预后。CIMP最近也在多形性胶质母细胞瘤中被研究，在这个研究中CIMP也被发现与较好的预后相关；比较于CIMP-negative样本，CIMP-positive样本在诊断时显著的比较年轻（平均年龄36 岁 vs 59 岁），与*IDH1*体细胞突变紧密相关，并且有显著好的生存时期。

甲基化也能够用作预测生物标记。例如，在角质母细胞瘤中，MGMT DNA修复基因的甲基化被报告对替莫唑咬有较好的应答相关及较好的临床结果相关。的确，MGMT启动子甲基化，出现在大约45%的样本中，被发现与来自替莫唑咬治疗的明显的效果相关，没有MGMT甲基化的患者中，替莫唑咬治疗效果明显减少。这些数据表明肿瘤甲基化谱能够有效用于风险分类和用于治疗的确定。

二、整合遗传与表观遗传特征识别疾病相关基因 》》》

经典的遗传与表观遗传是一个事物的两个方面，二者相互依存而又相互区别地构成一个整体，这样人类基因组就包含两个信息。遗传学提供合成生命所必需的蛋白质的模板，而表观遗传学的信息则提供了何时、何地以及如何应用上述遗传学信息的指令。整个基因组通过DNA精确地复制、转录、翻译，保证了遗传信息的稳定性和连续性，同时又通过表观遗传学机制，使基因组在内外环境条件下选择性地表达信息，最终形成遗传性状。

过去研究者们主要关注于癌症在遗传方面变异的研究，各种关于癌症的研究识别了大量的与癌症相关的遗传突变基因。随着对癌症研究的逐渐深入以及表观遗传学的发展，人们了解遗传突变和表观遗传变异共同导致了癌症的发生与发展。

Goh等人使用基因组特征对癌症的超甲基化基因进行排秩。他们使用计算的方法对实验证实的癌症超甲基化基因进行排秩。为了构建判别模型，他们同时选择基因组范围的基因作为对照，识别过甲基化基因作为潜在的诊断和预后的生物标记。

Yang等人开发了一个名为PGnet的算法，并构建了一个包含表型，基因和表观遗传调控相关的机制锚定网络。研究了132个（9种不同表型和三种对治疗响应的度量）急性淋巴细

胞白血病（ALL）微阵列。通过随机抽样可以从ALL表型得到一个稳定的子网。使用这个稳定的网络对ALL复发进行预测，结果评估是显著的（p=0.03）。DNA启动子区域的CpG岛甲基化是导致癌症中异常基因表达的一个机制。

Loss等人利用logistic回归模型构建基因启动子区域甲基化与表达之间的关系并对其排秩，在45个乳腺癌细胞系中识别了58个基因作为表观遗传调控的基因。Cui等人构建了一个手工获得的人类信号网络，将一组癌症突变基因和一组癌症相关的甲基化异常基因即应映射到信号网络中。研究发现癌症突变基因主要富集在正向信号调节通路中，而甲基化异常基因主要的富集在癌症细胞中负向调节通路中。

Liu等人利用网络理论并结合表观遗传和遗传特征，在癌症中识别甲基化异常的基因（见实例2）。他们整合文献证实的蛋白质互作数据和DNA甲基化数据，构建人类加权网络（WHPN）。WHPN呈现了基因对间的NDA甲基化水平的相互关系。在这个网络中，利用获得的癌症相关甲基化异常基因，获得一个与癌症紧密相关的子网络（CASN）。通过比较网络的拓扑特征，发现CASN有比WHPN更加紧密的网络结构。利用邻接加权规则在子网络识别了154个潜在的癌症相关的甲基化异常基因。发现这个基因主要参与调节细胞凋亡的生物学过程，并且其中很多基因在癌症会发生不同程度的差异表达。最后，通过结合手动确认的文本挖掘，发现这些识别的潜在癌症相关的甲基化异常基因中，一些能够作为癌症诊断和预后的生物标记，以及药物应答靶点。

第五节

人类表观基因组计划及意义

Section 5 The human epigenome project and its significance

一、表观基因组计划介绍

(一) 人类表观基因组计划(human epigenome project, HEP)

开始于1999年的第一个国际性表观基因组计划即人类表观基因组计划,目的是公布在人类健康组织和细胞中生成从识别出的可变甲基化位置而产生的人类DNA甲基化参考图谱。HEP利用高通量亚硫酸盐PCR Sanger 测序方法绘制了人类12个组织和细胞中3条染色体的DNA甲基化图谱。一些研究工作利用HEP的数据分析发现DNA甲基化有组织特异性、在进化保守区域频繁发生改变、个体发育过程中相对稳定等现象。在以后的研究中,为了完善HEP的数据, Down等人利用MeDIP-chip实验绘制了16个组织或细胞的DNA甲基化图谱, Rakyan等人使用MeDIP-seq 绘制了第一个人类全基因组的甲基化图谱。

(二) 癌症基因组图谱(the cancer genome atlas, TCGA)

由美国发起的TCGA计划是第一个从遗传学及表观遗传学两个方面的变化来考察人类癌症,目前已经收集了比较全面的胶质瘤、肺癌和卵巢癌的数据,其他癌症的数据正在不断更新,计划在今后的5年将公布超过20种肿瘤及成千个样本的遗传学及表观遗传学数据。TCGA提供了更深刻的理解人类癌症基因组及表观基因组的改变在癌症发生发展中的作用。

(三) NIH表观基因组路线图计划(RoadMap Initiative)

NIH表观基因组路线图的目的是以人类胚胎干细胞为中心开发广泛的表观基因组图谱以及开发新的工具用于分析这些数据,解决人类健康和疾病的表观遗传学研究中遇到的各种问题。该表观基因组学项目的目标是: ①创建国际化的表观基因组研究协会; ②开发标准的表观基因组学研究平台、软件以及研究标准; ③实施规范性规则评估表观基因组的改变; ④开发新的技术用于单细胞的表观基因组分析及活体组织表观遗传活性的图像; ⑤建立一个公共的数据资源加快表观基因组数据及方法的应用。这个计划已经成功的应用在单碱基水平胚胎干细胞和分化的成纤维细胞的甲基化图谱的绘制和分析。

(四) 国际人类表观基因组协会(international human epigenome consortium, IHEC)

IHEC是从2010年开始发起,目的是产生1000个人类和非人类的参考表观基因组图谱,

图谱广泛涉及各种组织和细胞、干细胞、环境和疾病(包括癌症)状态下表观基因组图谱的大量数据。该协会要建立一个表观基因组研究的公共交流平台,如培育表观基因组研究团体、组织各种学术会议、为低年资的研究者搭建研究平台以及为外界人士建立普及性网站等工作。

二、环境表观基因组与人类疾病

暴露在有害物质的环境里会导致表观遗传的改变,随着发育中哺乳动物表观基因组程序的变化,表观遗传进程适应包括环境、饮食甚至是行为等外在因素的变化。这种可塑性被认为是生物体迅速响应和适应外界刺激的反映,但也赋予生物体甚至是它们的后代记忆的能力(如果这样的刺激接触是在成年时接触到的)。最近的哺乳动物系统研究提供了这个与遗传的新达尔文理念相对的亚稳态的明确例子,特别是大量证据表明胎儿所处的环境影响表观遗传进程导致成长过程中对慢性疾病易感性的增加。这些结果对于人类健康有着重大影响且提供人类表观基因组图谱绘制的进一步的推动力。

双酚A诱导的胎儿表观基因组中低甲基化能够通过孕妇膳食补充而使甲基供体被消除,表明饮食的变化可以抵消环境毒物对于发育中的胎儿的潜在有害影响。通过怀孕和断奶期间改变孕妇营养的模仿人类病理的动物模型对于染色质的表观遗传修饰和后代表型的后续影响一直表现出连续的显著反应。例如,子宫胎盘功能不全能降低后代肾脏的DNMT1活性,造成*p53*启动子的低甲基化和*p53*表达的增加,有助于提高p53-介导的细胞凋亡从而降低肾小球数且引起高血压。产妇高脂肪的饮食有利于增加组蛋白H3乙酰化和降低组蛋白去乙酰化活性增加,和胎儿脂质代谢障碍代谢相关基因表达的增加相一致。这些和其余类似的研究明确表示产妇营养环境能显著影响胎儿表观基因组,直接有利于后代的健康。

尽管产妇营养对于子宫内胚胎发育的表观遗传学修饰有作用,母体的行为能改变新生儿的基因表达的表观遗传模式,一旦建立,会坚持到成年。综上,这些研究的结果导致了由生物体在发育的特别脆弱或是表观遗传不稳定时期由环境引起的表观遗传修饰参与人类疾病的病原学,在这些关键的发育窗口处这可能由膳食补充轻而易举地预防。

三、人类的表观基因组关联研究

人类复杂疾病的遗传和非遗传决定因素是生物学研究的一个重大挑战。最近几年,GWAS已经发现超过150种疾病的多于800个相关SNP和其他特征。虽然对于人类复杂疾病的完全遗传基础还是未知的,但重新排序外显子和最后的全基因组分析有助于识别大多数致病的遗传变异。然而,现在发现非遗传变异,包括表观遗传修饰的因素,能够对复杂疾病的病原学研究及发展提供帮助。因此在2008年表观基因组关联分析(epigenome-wide association study, EWAS)第一次被提出。由于DNA甲基化检测技术的成熟,使EWAS研究成为可能。

目前表观遗传学标记的完整图谱还不是很清楚,但是人们推测这个数据可能是巨大的,就二倍体人类表观基因组包括>108Cs(其中>107是CpGs)并且>108组蛋白尾部有可能发生共价修饰。在一个单独CpG位点上的DNAm变异,被认为是一个甲基化变异位点(methylation

variable positions, MVP),可以被认为是一个SNP的表观遗传学等价物。在复杂疾病中DNAm变异的角色,已经大体上在癌症中有了探索,可以被看作是早期EWAS研究。这些研究中发现,癌症发展和CGIs上DNAm的增加、印记缺失和表观遗传重复元件的重组有关,尤其是微卫星DNA上DNAm的缺失。糖尿病或者自身免疫性疾病中表观遗传学成分包括以下方面。第一,任何复杂疾病的同卵双胞胎的一致性是几乎从来没有100%。最近,系统性红斑狼疮和自闭症不一致的同卵双胞胎的小规模EWASs,已经发现同卵双胞胎内的疾病相关的表观遗传差异。第二,几个复杂疾病的发生率,例如1型糖尿病,在一般人群中上升,并且在流动人口中频繁变异,暗示了非遗传因素的一个角色。第三,流行病学证据说明一个次优的子宫内或者儿童早期环境,能够对成年期疾病的结果有影响(例如2型糖尿病),这个现象术语叫做“发育重编程(developmental reprogramming)”。当前,在子宫环境中分子标记主要的候选是表观遗传学修饰,包括DNAm。

潜在的基因型影响表观遗传变异,如同最近几个研究论证的。位点隐匿的遗传变异,影响甲基化状态,被称做甲基化数量性状位点(methQTLs)。在大多数methQTLs中,cis-genotype的相关是最显著的。在trans中,有一些证据表明,遗传变异也能够影响表观遗传状态,但是这似乎不像cis-effects那样普遍。要重点注意的是,大多数前人研究中,遗传变异的真正原因是未明确识别出来的,并且大多数meQTLs不是通过严格cis-genotype和epigenotype一对一的相关识别出来的;而是,一种特殊的基因形成一个甲基化增长的可能性。Feinberg和Irizarry最近已经讨论了在鼠和人类基因组中遗传变异的证据,不改变平均表型but rather表型的的变异性;这可能通过异变的甲基化区域(VMRs)被表观遗传学介导。MethQTL的存在为整合GWASs和EWASs,来发现基因型通过表观遗传变异发挥它们的功能、提供一个强有力的论据。

这些methQTLs也能够影响等位基因特异性甲基化(ASM)。关于这点,这种稳态的甲基化水平在相同细胞中两个等位基因上不同。然而,ASM也能够发生在一些缺乏任意特异的基因型——表观遗传型相关之中。例如,潜在的印记、X染色体失活和随机单等位基因甲基化是,不由甲基化和非甲基化等位基因之间潜在的基因型差异引起的ASM的全部实例。

因此一个细胞的表观遗传组是高度动态的,受一种复杂的遗传和环境相互影响所支配。正常细胞功能依赖于表观遗传自动调节功能维持,这在许多表观遗传扰动和人类疾病之间相关报道中突出显示,尤其是癌症。

第六节

应用实例：用于疾病的风险预测、诊断、预后及治疗的表观基因组数据分析

Section 6　Application examples: Analysis of epigenome data for risk prediction, diagnosis, prognosis and therapy of disease

经典的遗传学不足以解释表型的多变性，也不能解释相同的DNA序列、同卵双胞胎、克隆动物各自之间存在的表型差异和不同的患病情况。表观遗传学对这些现象提供了部分解释。DNA甲基化是目前人们了解最多的表观遗传标记。最开始发现了人类肿瘤中DNA的整体低甲基化，后来又发现了高甲基化的抑癌基因，最近，又发现由DNA甲基化导致的microRNA 基因的失活。

此外，最近的研究也发现了疾病状态下异常的组蛋白修饰模式，在人类肿瘤中，组蛋白H4修饰呈现单乙酰化和三甲基化的缺失，这些变化发生在早期且在肿瘤发生过程中不断积累，而且组蛋白修饰酶序列的表达量可以用于分类癌症。DNA甲基化和组蛋白修饰有潜在的临床用途。DNA高甲基化作为癌症一个标记，可以在所有类型的体液和活体标本中检测到。DNA甲基化和组蛋白修饰是可逆的，因此可以通过药物重新激活被沉默的抑癌基因或者抑制失控的致癌基因的表达。

有人提出应用与特定基因启动子结合的转录因子来开发表观特异的治疗药物。这些表观遗传调控机制的阐明及疾病诊断及治疗方面的尝试，促使人们更加深入地研究表观遗传修饰在预测、诊断、预后以及治疗疾病方面的作用。下面介绍这些研究中两个典型的实例，以期促进国内疾病表观遗传学的不断发展。

实例1　人类结肠癌甲基化组差异甲基化分析

人们已经知道癌症中异常的DNA甲基化的存在，包括致癌基因的低甲基化和肿瘤抑制基因的高甲基化。然而，大多数癌症甲基化研究认为功能上重要的甲基化发生在启动子区域，癌症中的DNA甲基化改变多发生在CpG岛。但还没有理解全基因组癌症和正常分化中DNA甲基化缺失和获得的关系。在这些实验中，人们关注三个关键的问题。第一，不同组织类型的DNA甲基化差异发生在哪儿？该研究采用一个全基因组的方法-基于芯片的高通量的相对甲基化分析（comprehensive high-throughput array-based relative methylation,

CHARM），从5个实验（个体）中选用三个正常组织类型代表三个胚层—肝（内胚层），脾脏（中胚层），脑（外胚层）。不同于之前组织中甲基化的研究，该研究是全基因组的，而且组织来自相同的个体，可以控制个体间差异的可能性。第二，癌症中DNA甲基化改变发生在哪儿？低甲基化和高甲基化之间的平衡是什么？基于这个目的，该研究检验13个结肠癌和相应的正常黏膜。第三，这些甲基化变化的功能作用是什么？为此，该研究进行小鼠组织甲基化和表达分析的比较表观基因组研究。

该研究通过比较每对组织平均M值的差异（ΔM）定量来表示差异甲基化，并发现了识别出16 379个组织差异甲基化区域（T-DMRs），大约76%的T-DMRs定位于CpG岛区域2kb的位置，定义为CpG岛边缘；另外分析同一个体的13个结肠癌和匹配的正常黏膜的DNA甲基化，识别了2707个在癌症中显示差异甲基化的区域（C-DMRs），且发现多数结肠癌甲基化改变不发生在启动子区域，也不在CpG岛，而也在CpG岛边缘（图11-12）。

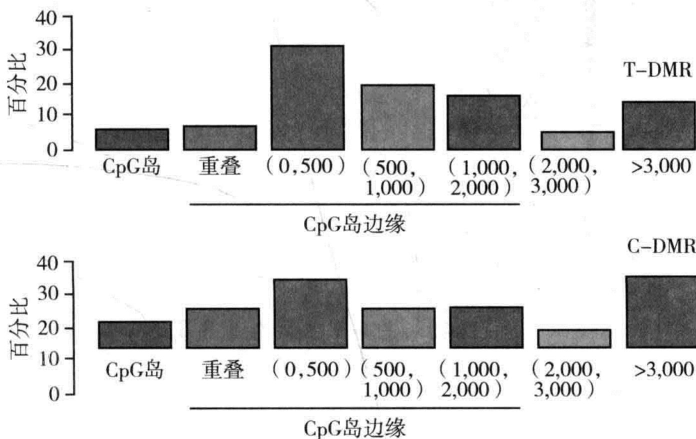


图11-12　T-DMR和C-DMR与CpG岛和CpG岛边缘相对位置的分布

来源于：Irizarry RA .The human colon cancer methy lome shows similar hypo-and hypermethy lation at concerved tissue-specific Gp G island shores.（2009）Nat Genet.2009；41：178-186.

该研究还探讨这些差异甲基化和相关基因表达的功能关联。为了研究组织和癌症差异甲基化，研究者分析了来自相同样本的5个大脑和肝脏的和4个结肠癌和正常黏膜组织的基因表达。所有的样本来自基因组甲基化分析数据。T-DMRs甲基化和差异基因表达显示强的逆相关（图11-13），即使这些DMRs不在CpG岛而是在CpG岛边缘。尽管C-DMRs的数量比T-DMRs较少，但也显示了与基因表达差异的显著关联。这些基因表达和CpG岛边缘甲基化之间的功能关联暗示CpG岛边缘具有调控功能的可能性。

由于C-DMRs和T-DMRs都定位于CpG岛shores，研究者又研究了它们是否定位于相同的位置。惊奇地发现C-DMRs与T-DMRs存在显著地位置重叠（45%~65%）。C-DMRs和T-DMRs的关联是如此显著，用C-DMRs的平均M值对大脑，肝脏和脾脏进行非监督聚类，能很好区分这些组织（图11-14）。GO功能富集分析发现C-DMRs富集发育和多能性相关的基因。因此癌症特异DNA甲基化主要涉及组织间DNA甲基化变化的相同位点，尤其是与发育相关的基因，这一结果表明表观遗传改变影响组织特异分化是表观遗传改变引发癌症的主要机制。

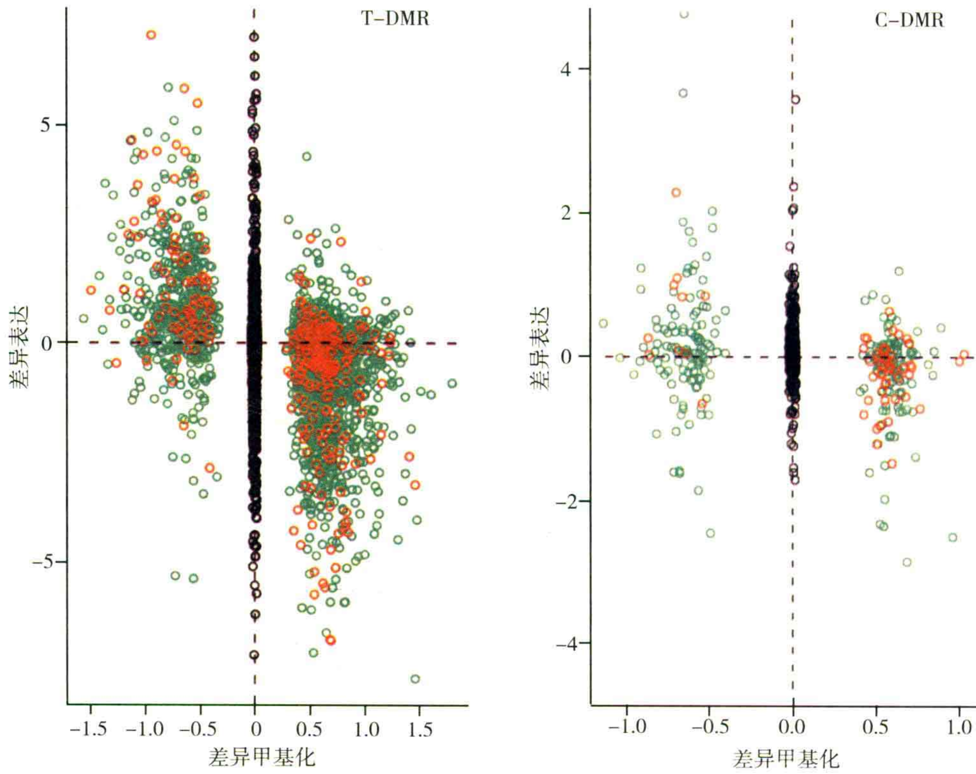


图11-13 差异甲基化与差异表达

来源于：Irizarry RA .The human colon cancer methy lome shows similar hypo-and hypermethy lation at concerved tissue-specific Gp G island shores.(2009)Nat Genet.2009；41：178-186.

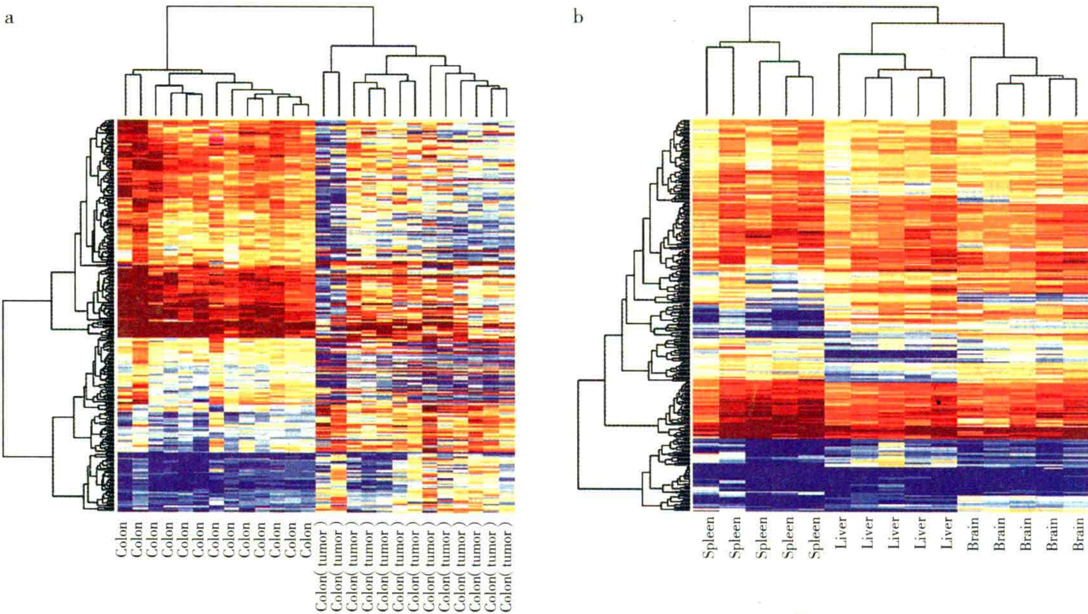


图11-14 C-DMR甲基化正确分类结肠癌以及正常组织

来源于：Irizarry RA .The human colon cancer methy lome shows similar hypo-and hypermethy lation at concerved tissue-specific Gp G island shores.(2009)Nat Genet.2009；41：178-186.

实例2 基于加权蛋白质互作网络优选癌症相关甲基化异常基因

网络理论为系统的研究疾病提供了一个便利的平台。基因间复杂的关系能够通过网络理论清晰的呈现出来。基于网络理论,结合表观遗传特征和遗传特征,识别癌症中发生甲基化异常的基因,可能作为癌症诊断和预后的生物学标记。该研究DNA甲基化以及蛋白质互作构建了加权蛋白质互作网络,并获取癌症相关子网络,识别潜在的癌症相关的甲基化异常基因。在这个研究中,人们同样关注三个主要的问题。第一,构建的加权蛋白质互作网络以及癌症相关子网络具有何种特点,以及它们之间有何差异。癌症相关子网络是否具有实际的意义。第二,优选的基因具有何种生物学功能,是否能够在癌症中发生甲基化的异常,并影响癌症的发生及发展。第三,优选的基因是否能够作为癌症诊断和预后的生物学标记,以及药物作用反应靶点。该研究针对以上问题进行了分析。

该研究利用DNA甲基化特征和蛋白质互作特征构建了加权人类蛋白质互作网络(WHPN),并通过种子基因集在WHPN中获得了一个癌症相关的子网络(CASN)(图11-15)。这两个网络的拓扑特征被比较分析,包括网络的度,聚类系数和平均路径长度,结果表明CASN有比WHPN更加紧密的网络结构,并显著于随机情况。通常认为在对相似疾病表型有贡献的突变的蛋白质之间经常发生直接的或者间接地互作。癌症被认为是一些相关通路的失调引起的结果,因此如果一个基因在 PPI 网络中与癌症基因接近,则认为它更有可能参与癌症的一系列事件。而子网络中的基因都是与甲基化异常癌症基因相关的基因,这些基因与种子基因一样有可能参与相同或相似的生物学过程,因而在癌症中可能同样发生甲基化水平的改变。

该研究利用邻接加权规则在CASN网络中识别了154个潜在的癌症相关甲基化异常基因。这些基因主要参与调控有关调控细胞凋亡及程序性死亡的生物学过程和p53 signaling pathway和Wnt signaling pathway等癌症相关通路。这些结果表明这些基因可能在癌症相关的

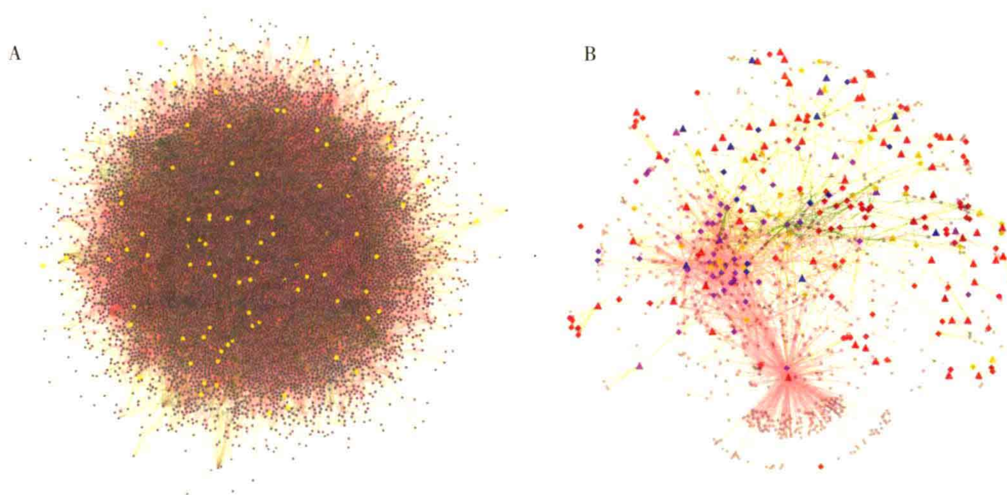


图11-15 WHPN网络和CASN网络

来源于: Liu H .Priontizing cancer-related genes with aberrant methylation based on a weighted protein-protein interaction network.BMC Syst Biol.2011:155-158.

生物学过程及通路中发生异常，产生甲基化水平的异常变化从而影响癌症的发生以及发展。通过对相关癌症表达谱的SAM（significant analysis of microarrays）分析发现，优选基因集中的一部分基因能够在相应癌症中发生表达的异常。

该研究通过搜索PubMed评估了优选基因与癌症的关系。结果发现其中43个基因在发表的文献中与各种癌症以及甲基化异常相关，其中10个基因被报道在癌症中会发生甲基化的异常现象（表11-5）。然后通过对PubMed自动化的搜索以及手工校对，发现27个优选基因能够作为癌症以及其他复杂疾病的诊断标记，20个优选基因能够作为预后标记。最终，将优选基因匹配DrugBank药物标点列表，发现31个优选基因能够作为药物反应的应答靶点。这些结果表明优选基因可能在癌症中发生甲基化异常，导致基因表达的活化或抑制，作用于癌症的发生以及发展，是癌症潜在的诊断和预后标记。

表11-5　通过PubMed文献证实的疾病诊断预后标记基因及潜在的药物靶点基因

基因名	Entrez基因编号	诊断标记	预后标记	药物靶点
CREBBP	1387	▲		⊙
EP300	2033	▲	★	
HIF1A	3091		★	⊙
PRMT1	3276	▲	★	⊙
PML	5371		★	
med1	5469			⊙
tp63	8626	▲	★	
PRKCDBP	112 464	▲		
MANEAL	149 175			
Rasef	158 158			

注：数据来源：Liu H .Priontizing cancer-related genes with aberrant methylation based on a weighted protein-protein interaction network.BMC Syst Biol.2011:155-158.

（张　岩　刘洪波）

参考文献

1. Kelly TK, De Carvalho DD, Jones PA. Epigenetic modifications as therapeutic targets. *Nat Biotechnol* 2010, 28 : 1069-1078.
2. Jones PA, Baylin SB. The epigenomics of cancer. *Cell* 2007, 128 : 683-692.
3. Maunakea AK, Chepelev I, Zhao K. Epigenome mapping in normal and disease States. *Circ Res* 2010, 107 : 327-339.
4. Lv J, Liu H, Su J, et al. DiseaseMeth: a human disease methylation database. *Nucleic Acids Res* 2012, 40 : D1030-1035.
5. Zhang Y, Lv J, Liu H, et al. HHMD: the human histone modification database. *Nucleic Acids Res* 2010, 38 : D149-154.

6. Rakyan VK, Down TA, Balding DJ, et al. Epigenome-wide association studies for common human diseases. *Nat Rev Genet* 2011,12 : 529-541.

7. Portela A, Esteller M. Epigenetic modifications and human disease. *Nat Biotechnol* 2010,28 : 1057-1068.

8. Koga Y, Pelizzola M, Cheng E, et al. Genome-wide screen of promoter methylation identifies novel markers in melanoma. *Genome Res* 2009,19 : 1462-1470.

9. Irizarry RA, Ladd-Acosta C, Wen B, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* 2009,41 : 178-186.

10. Zhou VW, Goren A, Bernstein BE. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* 2011,12 : 7-18.

11. Hansen KD, Timp W, Bravo HC, et al. Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* 2011. DOI: 10. 1038/ng. 865.

12. Zhang Y, Liu H, Lv J, et al. QDMR: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic Acids Res* 2011,39 : e58.

13. Barski A, Cuddapah S, Cui K, et al. High-resolution profiling of histone methylations in the human genome. *Cell* 2007,129 : 823-837.

14. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 2011,12 : 56-68.

15. Mohammad HP, Baylin SB. Linking cell signaling and the epigenetic machinery. *Nat Biotechnol* 2010,28 : 1033-1038.

16. Liu H, Su J, Li J, et al. Prioritizing cancer-related genes with aberrant methylation based on a weighted protein-protein interaction network. *BMC Syst Biol* 2011,5 : 158.

17. Laird, PW. Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet* 2010,11 : 191-203.

第十二章
CHAPTER 12

药物生物信息学

PHARMACOBIOINFORMATICS: REVOLUTIONIZING DRUG DISCOVERY RESEARCH

传统的药物研发模式盲目，耗时耗资巨大，成功率低，甚至有些药物上市后由于严重不良反应而惨遭淘汰。随着基因组学、转录组学、蛋白组学、代谢组学和变异组学等组学的蓬勃发展，产生了大量的高通量数据资源以及相应的分析技术，其成果不仅给人类认识自身本质和疾病发生机制提供了新的机遇，同时还为制药工业带来了前所未有的机会和全新的药物研发理念。

在药物的研发过程中，药物生物信息学的技术和手段可以明确靶点、早期预测药物的成药性，并预测药物适合人群，体现了药物生物信息学在药物研发领域的巨大优势，使之成为后基因组时代最引人瞩目、发展最迅速的学科之一。2003年美国FDA发行了“Draft guidance for Industry Pharmacogenomic Data Submission”，我国也在2007年首次把药物生物信息学研究计划纳入了863项目。

药物生物信息学是一个将信息学和计算机科学的原理和技术应用于药物发现和药物防治的一门新学科。它整合了多学科的原理和方法如生物信息学、化学信息学、化学基因组学以及其他发展完善的学科，如药理学、药物化学、理论化学和药学实践等，从系统和全局的角度为制药行业提供理论与技术工具，这是传统药学无法比拟的。

尽管药物生物信息学在药物研发的各个阶段体现了极大优势，但由于本书篇幅所限，本章只向大家介绍三方面内容：药学相关信息资源；生物信息学方法鉴别和验证药物靶标；个性化给药基础——药物基因组学。

第一节

药学相关信息资源

Section 1 Bioinformatics Resources of Pharmaceuticals

一、药物综合信息数据库

DrugBank（http://www.drugbank.ca/）为综合性免费药物信息资源，覆盖大量药物及其靶标相关信息。截止到2012年3月，数据库中包含有6711条药物的相关条目信息，其中包括FDA批准的1441种小分子药物、134种蛋白质和多肽类生物技术药物、84种营养制品和5084种处于实验研究阶段的药物。除此之外，数据库还提供了4231条非冗余的与药物相关的蛋白质序列信息（如药物的作用靶标、代谢酶、转运蛋白、载体）。每种药物提供了150多项信息，包括药物名称、化学结构、蛋白和DNA序列以及药理学、药物经济学、药物间相互作用、靶点、相关的酶、单核苷酸多态性、药物副反应和相关文献的链接等。

DrugBank数据库的设计者之一，David Wishart博士认为DrugBank数据库是目前唯一一个融合了生物信息学和化学信息学知识的数据库，是将药物的详细信息与药物广泛作用的靶标信息的完美结合。

DrugBank数据库提供多种浏览、查询的方式，便于访问者对相关知识进行查询和了解。图12-1是DrugBank数据库的主界面。

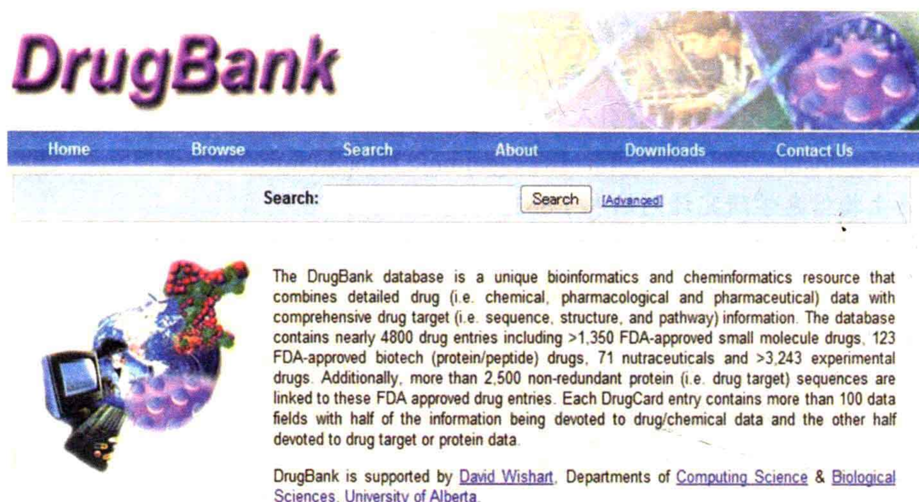


图12-1 Drugbank主页界面

二、与药物靶点发现相关数据库

(一) 疾病相关的基因数据库

目前,已有一些数据库存储了与疾病相关的基因信息,方便研究人员对相关基因或蛋白质进行查询和比较。与人类疾病相关的基因以及基因敲除时的异常情况存储在online mendelian inheritance in man(OMIM, http://www.ncbi.nlm.nih.gov/omim/), entrez gene(http://www.ncbi.nlm.nih.gov/gene/), the human gene mutation(http://www.hgmd.cf.ac.uk/ac/index.php), catalogue of somatic mutations in cancer(COSMIC, http://www.sanger.ac.uk/genetics/CGP/cosmic/)和cancer gene census(http://www.sanger.ac.uk/genetics/CGP/Census/)等数据库中。其中, OMIM数据库是分子遗传学领域最重要的生物信息学数据库之一,是人类基因和遗传性疾病的电子目录,提供疾病与基因、文献、序列记录、染色体定位及相关数据库的链接等内容。该数据库可以通过ENTREZ进行搜索,并且利用“limit”选项限制所搜索的染色体位置或类别等内容。COSMIC数据库存储了癌症相关的候选基因,提供体内基因敲除信息以及人类癌症的相关细节。cancer gene census数据库对癌症相关的基因进行分类,这些基因在敲除时与癌症表现出可能的因果关联。其中的GeneRif系统可以提供与疾病高度相关基因的注释信息。此外,基因组规模的关联数据库、遗传关联数据库和小鼠基因敲除数据库等也为基因查询提供了丰富的注释信息。

(二) 疾病相关的基因芯片数据库

基因芯片数据是药物靶标发现的重要来源,人们已经建立了一些专门的数据库用于存储疾病相关的基因芯片数据。gene expression omnibus(GEO, http://www.ncbi.nlm.nih.gov/geo/)作为存储基因芯片的主要数据库资源,包含了丰富的癌症相关的基因芯片数据。当以“Homosapiens”和“Cancer”作为关键词进行查询时,返回了350个数据集。2003年10月, Daniel 等建立了ONCOMINE 数据库(http://www.oncomine.org/),专门收集癌症相关的基因芯片数据集,提供在线的数据挖掘和基因组规模的表达分析。在ONCOMINE 数据库的三个版本中,包含了264个基因表达数据集,超过2万个癌症组织和正常组织的样本数据。其他基因芯片数据库包括斯坦福基因芯片数据库(http://genome-www5.stanford.edu/MicroArray/SMD/)、EBI芯片表达数据库(http://www.ebi.ac.uk/arrayexpress/),以及MIT 癌症基因组工程(http://www.broad.mit.edu/cancer/)等。

(三) 候选靶点信息资源

治疗靶点数据库(therapeutic target database, TTD, http://bidd.nus.edu.sg/group/cjttd/)是一个免费的数据库,覆盖了2025个药物靶点(包括364个确认靶点,286个试验阶段靶点及1331个研究靶点)及靶点相关疾病和信号通路,同时包含17 816个药物/配体(包括FDA认可药物1540种,临床试验阶段药物1423种,实验研究阶段药物14 853种;其中有小分子药物14 170种,反义核酸类药物652种),同时提供链接到其他数据库,以方便检索蛋白功能、氨基酸序列、三维结构信息、配体结合特性、药物结构、治疗应用等内容。TTD虽然设计目的是专门化数据库,但是实际上也是一个关于药物的综合性数据库。

non-redundant databases（NRDB，http：//linux.mlst.net/nrdb/nrdb.htm/）由NCBI建立，数据来自genpept（genBank CDS自动翻译的数据库）、PDB序列数据库、SWISS-PROT数据库等，是比较完全且包含最新信息的蛋白质数据库，是检索药物靶点的主要信息来源。实际上NRDB中仍然有一些冗余信息。另外，NRDB数据库也被作为NCBI提供的BLAST算法搜索服务时检索的默认数据库。

潜在药物靶点数据库（potential drug target database，PDTD，http：//www.dddc.ac.cn/pdtd/）为国内建立的免费药物靶点数据库，收集了已知的和潜在的药物靶点的三维结构数据，是反向对接筛选候选药物靶点软件常选用的数据库之一。此数据库目前包括1207个晶体结构数据，涵盖了841种不同药物靶点。这些靶点按照治疗应用领域和靶点的生物化学性质分成十多类，支持多种检索方式，并可链接到其他数据库。

蛋白质信息学资源（protein-informatics-resource，PIR，http：//pir.georgetown.edu/），该数据库提供通用的蛋白质序列和功能数据，并链接到UniProtKB（http：//www.uniprot.org/help/uniprotkb）等多个数据库。此数据库的特色是可提供详细全面的蛋白质功能分类数据，其中包含药物靶点蛋白质的数据。

（四）与发现药靶相关生物学功能数据库

药物靶标通常具有特定的生物学功能，分析基因的分子类型（例如酶）、亚细胞定位（例如细胞表面）和生物学通路（例如血管新生）对于预测潜在药靶具有重要意义。基因本体论gene ontology（GO，http：//www.geneontology.org/）和京都基因与基因组百科全书数据库kyoto encyclopedia of genesand genomes pathways（KEGG，http：//www.genome.ad.jp/kegg/）提供了多个物种中基因的生物学功能、定位和通路信息。同时，有关蛋白质相互作用网络和生物学通路的数据库资源非常丰富，如database of interacting proteins（DIP，http：//dip.doe-mbi.ucla.edu/dip/Main.cgi/）、reactome（http：//www.reactome.org/ReactomeGWT/entrypoint.html/）、nature pathway interaction database（NCI，http：//pid.nci.nih.gov/）、human protein reference database（HPRD，http：//hprd.org/）和Biocarta（http：//www.biocarta.com/）等，更多的数据库列表可以参考http：//www.pathguide.org/。此外，有些数据库专门存储生物学网络的定量数据资源，例如bioModels（http：//www.ebi.ac.uk/biomodels-main/）和JWS online（http：//jjj.biochem.sun.ac.za/database/）数据库，收集了各种化学反应网络的数学模型，并且规模一直在稳步增加。

（五）不良反应数据库

Side effect resource（SIDER，http：//sideeffects.embl.de/）为药物不良反应数据库，包含了药物和副反应的相关信息。该数据库包含了888种药物的相关条目信息（798种FDA批准的药物和90种非FDA批准的药物）、1450种不同的不良反应和62 269个药物-不良反应关系对。

数据库采用文本挖掘的方法对多种资源的数据进行整理，同时给出了药物-不良反应对的不良反应发生频率信息。数据库将不良反应的发生率分成以下几个等级：postmarketing 、rare（<0.1%）、infrequent（0.1% to 1%）、frequent（1% to 100%）。

用户可从数据库中获得以下信息：药物相关的不良反应、药物化学结构、部分药物的靶标、可产生同种不良反应的药物、具有相同不良反应特征的药物、人群服用药物产生不良反应的概率以及施用安慰剂作为对照组时产生该不良反应的概率等。

同时，从search tool for interactions of chemicals（STITCH，http: //stitch.embl.de/）、PubChem（http: //pubchem.ncbi.nlm.nih.gov/）、Wikipedia 和Medpedia中都可以获得不良反应以及药物相关药理作用的描述信息。图12-2是SIDER数据库的主界面。

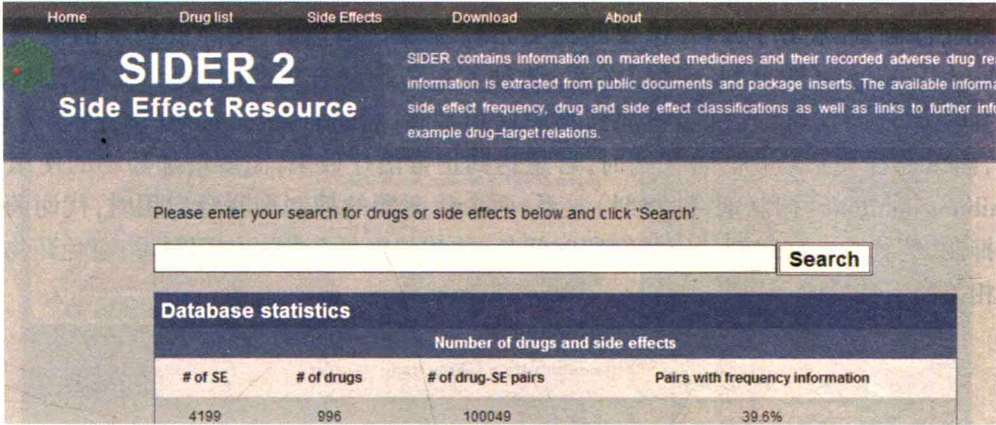


图12-2 SIDER主页界面

（六）药物相互作用查询数据库

Cytochrome P450 database（SuperCYP，http: //bioinformatics.charite.de/supercyp/），是基于细胞色素P450酶来分析药物间的相互作用的数据库。该数据库收录了1170种药物的信息、2785种基于细胞色素P450酶产生的药物相互作用、57种人类CYP家族酶的信息和1200个等位基因信息。对具有相同代谢酶的药物，数据库会提供出同类替代药的建议。图12-3为数据库的主页面。

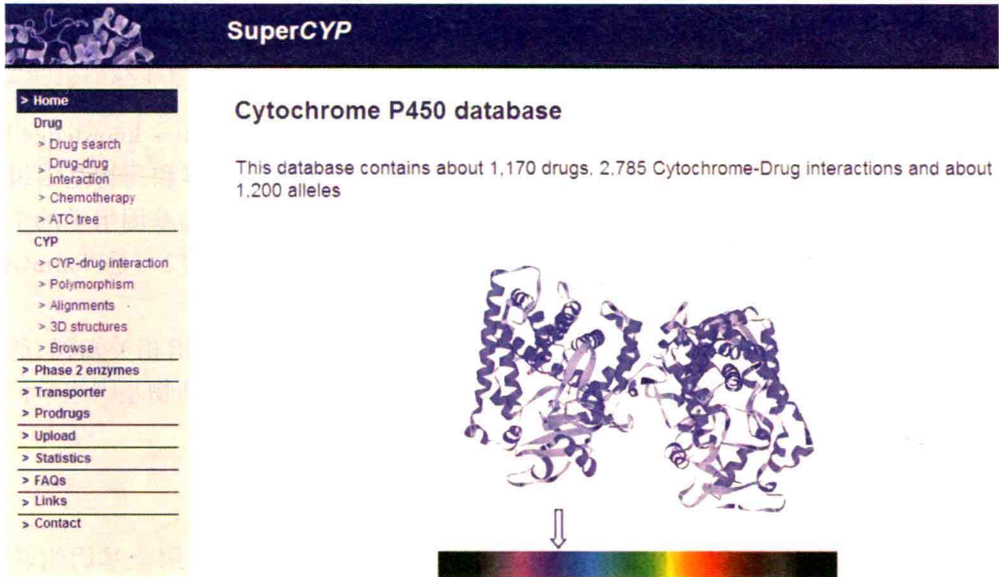


图12-3 SuperCYP数据库主页

数据库对基于CYP450产生的药物相互作用分为五种情况，并给出了相关的用药建议。具体分类情况如图12-4所示：①substrate-substrate（底物-底物）关系，即如果多种同为CYP同一代谢酶底物的药物联合应用时，很可能由于它们之间对酶的竞争使得药物代谢被抑制，药物会在组织器官中停留更长时间，故建议需要降低药物使用的剂量；②inhibitor-substrate（抑制剂-底物）关系，建议此种关系药物联合用药时，要降低药物的使用剂量；③inducer-substrate（诱导剂-底物）关系，此种关系药物联合用药时，因为代谢酶被激活，药物排出体外会加快，故建议增加药物的使用剂量；④inducer-inducer（诱导剂-诱导剂）关系，建议此种关系药物联合用药时，若想达到正常治疗效果，需要增加药物的使用剂量；⑤inhibitor-inhibitor（抑制剂-抑制剂）关系，两个代谢酶的抑制剂联合应用时，代谢酶的活性被抑制，药物在体内组织器官的存留时间长，故建议降低药物的使用剂量，避免药物蓄积产生相关的不良反应。

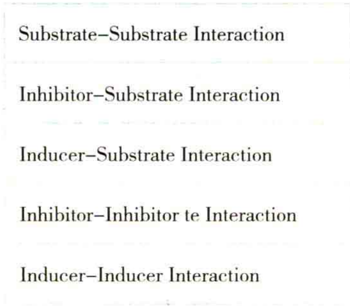


图12-4 SuperCYP数据库中药物相互作用情况分类

SuperCYP数据库中分别用s标识substrate（底物），inh标识inhibitor（抑制剂）和ind标识inducer（诱导剂）。在查询结果中，点击相关的标识会直接给出PUBMED的文献来源链接，方便访问者进行查询。在给出相关替代药物时，标志为绿色的药物是可供选择的替代药物。

（七）药物基因组学数据库

药物基因组学数据库（the pharmacogenetics and pharmacogenomics knowledge base，PharmGKB，http://www.pharmgkb.org/），是基于遗传药理学、药物遗传学和药物基因组学知识所建立起来的，主要是用来收集、分析、记录和传播遗传药理学和药物基因组学的主要数据和相关知识的数据库，所以它是药物基因组学的代表性数据库。图12-5是PharmGKB的主页界面。

PharmGKB数据库从单核苷酸基因多态性（SNP）、临床的药物基因组相关知识、通路信息、药物和小分子信息、药物基因组学相关的基因信息和与疾病相关的药物基因组学信息这六方面的内容对药物基因组学进行了描述。

（八）疾病-基因-药物联通图

The connectivity map（CMAP，http://www.broad.mit.edu/cmap/），是应用全基因组转录谱系统全面地描述生理、疾病和药物诱导等生物学状态，以GSEA（gene set enrichment analysis）算法提取并比较这些生物学状态的基因表达标识，将相同（似）或相反功能药物、药物适用疾病、药物作用途径（基因、通路）联系起来。

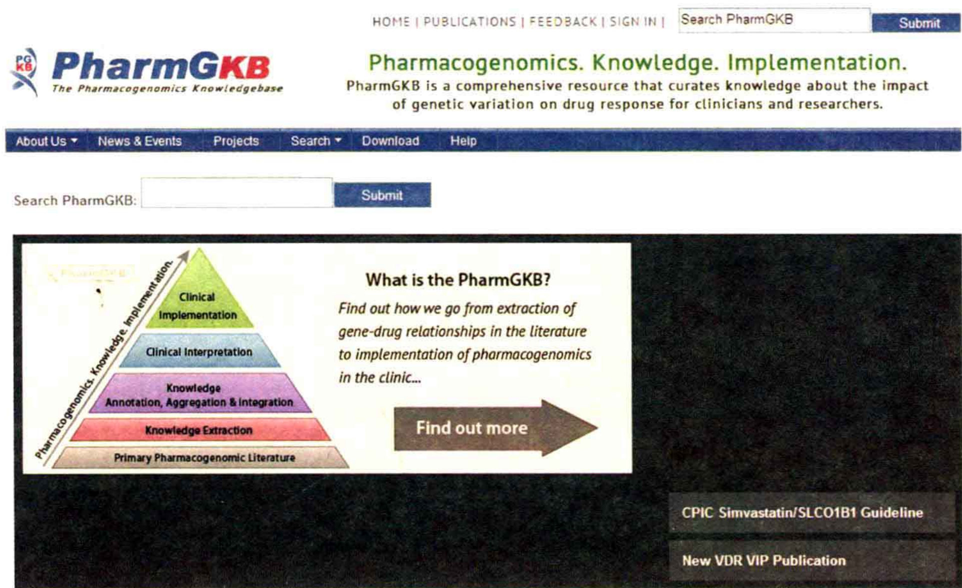


图12-5　PharmGKB数据库的主页界面

CMAP数据库的构建是基于联通图的基本假设，假设药物刺激前后全基因组范围的基因表达都将会发生改变，这种改变能够从本质上反映生物体系统对药物的应答情况即对药物的反应状态，从而可以代表药物的药理活性特征。这种描述药物特征的策略从药物作用本质出发，关注分子水平的变化，基于全基因组水平，因此很具有系统性。

图12-6举例描述了联通图的基本原理。用基因芯片检测某未知功能化合物刺激前后生物体（人、鼠、细胞系）的全基因组表达水平，比较刺激前后的表达谱水平，采用生物信息学方法提取基因组标识。然后，将差异表达基因标识录入到数据库的查询界面中，数据库采用GSEA算法评价此标识与库中药物标识的相似性，输出与已知药物功能相关联的药物，作用越相似的排序越靠前，作用越相反的排序越靠后，从而推知已知药物的未知功能。在图12-6中，绿色区域越上端的药物与已知药物的作用机制等越相近，红色区域越下端的药物与已知药物作用机制越相反。

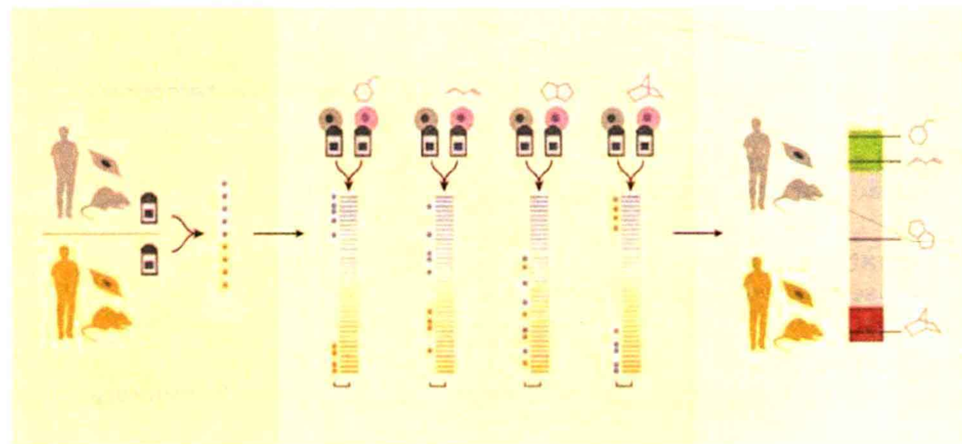


图12-6　联通图的基本原理

CMAP数据库主要用于发现共享作用机制和生化进程的小分子、疾病和药物之间的关联性，可以用于确认药物的作用机制、发现已上市药物的其他作用、预测未知作用机制药物的机制、研究疾病的生理机制等。

图12-7是CMAP数据库的登陆主界面，该数据库提供相关数据信息和查询结果的免费下载，但是需要注册后登陆。将注册的用户名和密码输入后进入数据库主界面，见图12-8。

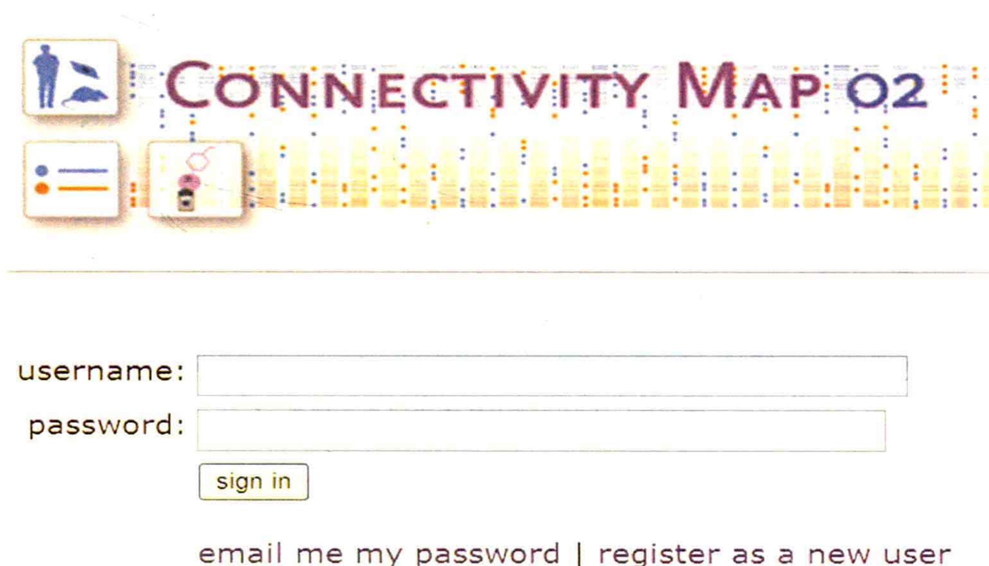


图12-7 CMAP数据库的登陆主界面

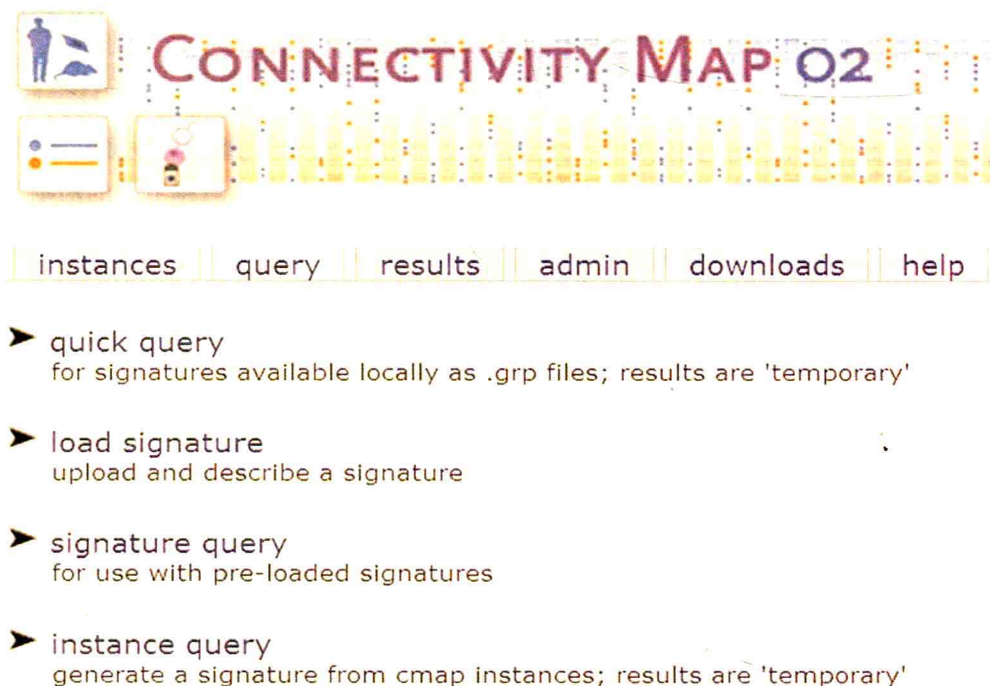


图12-8 CMAP数据库的主界面

第二节
基于生物信息学方法发现潜在药物靶标

Section 2 Bioinformatics-based approach to identify potential drug targets

建立系统、高效的药物研发创新体系是后基因组时代具有挑战意义的任务。发现并验证药物新靶标是研发创新药物的首要工作，人类基因组计划的完成推动了基因组学，蛋白组学和代谢组学的发展，为寻找药物靶标带来了新的机遇，而在高通量数据分析方面显示巨大优势的生物信息学技术已成为从庞大的组学数据中挖掘药物靶标信息的一种重要手段。

在药物靶标发现的过程中，生物信息学方法发挥了不可替代的重要的作用，尤其适用于大规模多组学数据的分析。目前，已涌现了许多与疾病相关的数据库资源，基于生物网络特征、生物芯片、蛋白质组、代谢组数据等建立了多种生物信息学方法发现潜在的药物靶标，并预测靶标可药性和药物副作用。

药物靶标是指体内具有药效功能并能被药物作用的生物大分子，如某些蛋白质和核酸等生物大分子。事先确定靶向特定疾病有关的靶标分子是现代新药开发的基础。在药物发现的漫长过程中，药靶发现是非常重要的一个限速步骤。药靶筛选和功能研究是发现特异的高效、低毒性药物的前提。

靶标发现与确证的一般流程（图12-9）是：利用基因组学、蛋白质组学以及生物芯片技术等获取疾病相关的生物分子信息，并进行生物信息学分析；然后对相关的生物分子进行功

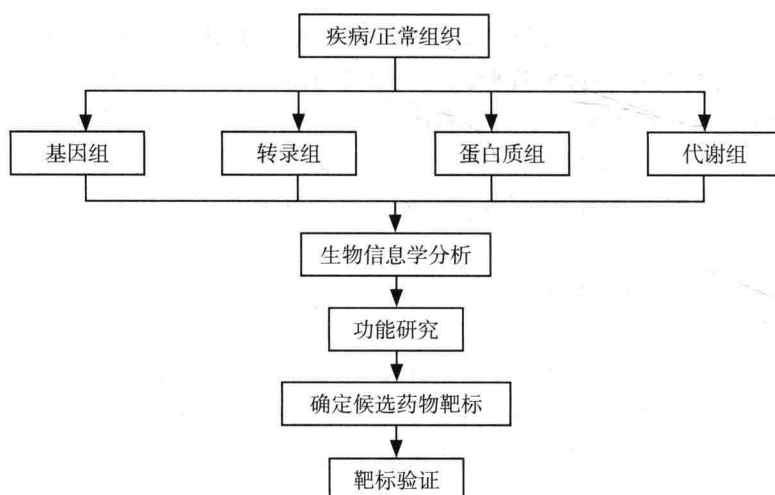


图12-9 **药物靶标发现一般流程**

能研究，以确定候选药物作用靶标；针对候选药物作用靶标，设计小分子化合物，在分子、细胞和整体动物水平上进行药理学研究；验证靶标的有效性。

常见的用于药靶发现的实验方法包括：微生物基因组学、差异蛋白质组学、磁共振（MR）技术、细胞芯片技术、RNAi技术、基因转染技术和基因敲除动物等。随着组学数据的积累，仅凭实验方法已经不能满足高通量大规模数据分析的需求。

在药物研发过程中，生物信息学方法对于相关数据的存储、分析和处理，以及如何有效地发现和验证新的药靶，发挥了重要的作用。

本节主要基于疾病相关的基因数据库、候选药靶数据库和基因芯片数据库等资源基础上，讨论基于多种组学数据进行药物靶标发现的生物信息学方法，如基于基因组、基因表达谱、蛋白质组、代谢组的方法以及整合多组学数据的系统生物学方法，最后描述生物信息学方法在药物靶标验证方面的应用，主要是预测蛋白质可药性以及药物副作用。

一、用于药靶发现的生物信息学方法

（一）基因组方法

丰富的基因组学数据为药靶发现提供了基础，目前已有多种方法可用于寻找新的药物靶标。其中，最常用的方法是同源搜索，采用序列比对软件寻找候选基因与已知癌症基因之间的序列同源性，如BLAST或基于隐马尔科夫的HMMER软件包等。然而，新的靶标与已知癌症基因的序列可能并不相似。因此，有必要分析已知药靶中更为普遍的结构特征，如信号肽、跨膜结构域或蛋白激酶域，此类生物信息学工具包括预测信号肽的SignalP和预测跨膜结构域的TMHMM。此外，还可以使用基因预测程序从人类基因组序列中预测新基因，寻找全新的药物靶标，常用的程序是 Genescan和Grail。

通过单基因敲除实验能够发现生物体中的必要基因（essential gene）。但以必要基因作为癌症治疗的靶标不仅能杀死癌细胞，对于健康细胞也可能是致命的。因此，大多数以单基因作为靶标的药物治疗是失败的，双基因的合成致死性（synthetic lethal）为抗癌药物的研究提供了新的前景。给定一个癌症相关的基因，如果该基因在癌细胞中功能缺失或者功能降低，那么以它的合成致死对象作为药靶就能构成肿瘤细胞的致死条件，同时降低对健康细胞的损伤。目前，仅在酵母中通过大规模的实验建立了全基因组的合成致死网络。通过同源预测等方法，Conde-Pueyo等重建了人的基因合成致死网络，为抗癌研究中候选基因靶标的筛选提供依据。目前已知的单基因病种类较少，仅限于基因组方法得到的药物靶标作用效果往往不够理想。随着后基因组时代的到来，其他组学数据在药物靶标发现中发挥了越来越重要的作用。

（二）基因芯片方法

基因芯片技术指将大量（通常每平方厘米点阵密度高于 400）探针分子固定于支持物上与标记的样品分子进行杂交，检测每个探针分子的杂交信号强度，进而获取样品分子的数量和序列信息。由于基因芯片技术的高通量、快速、平行化等特点，使得疾病相关的基因芯片数据资源非常丰富，利用基因芯片数据挖掘潜在药物靶标成为一种重要的途径。例如，在GEO数据库的基础上，Hu等建立了大规模的疾病-药物对应网络，帮助有效地识别药物靶标。

但由于基因芯片本身存在重复性较差和数据质量不高等问题，需要发展多种有效的分析方法，尤其是能够处理多个数据集、对噪声不敏感的统计方法，以提取海量数据中蕴含的有用信息。

1. 基于比较基因芯片数据　基因芯片能够一次性地记录疾病状态下成千上万个基因的变化情况。通过比较疾病组与正常组的基因芯片数据，寻找显著差异的基因集合，可用于预测相关的生物标志物或药物靶标。其中，寻找差异表达基因的计算方法很多，最直接的方法是测量变化倍数，即计算两个样本之间同一个基因的表达量之比。尽管变化倍数方法直观有效，但是该方法没有考虑噪声和生物学可变性，尤其是癌症这种本质上多相异质的复杂疾病。因此，更加通用的办法是采用尽可能多的疾病样本进行统计学分析，如ANOVA和 T-like 检验等。由于单个基因难以检测疾病状态下翻译模型的变化，生物标志物通常包括一组基因，需要一定的聚类方法寻找相关基因的组合。如GSEA方法能够评估两种生物学状态下一组基因集合的统计显著性，已广泛地应用于基因芯片数据的分析。

2. 多种来源的基因芯片数据的整合　由于单个芯片数据本身存在的噪声及系统偏差，预测结果往往存在误差。因此，最新的研究通过整合不同实验来源的多组基因芯片的数据，减少单个芯片实验中的误差影响，寻找更加通用的生物标志物和药物靶标。数据整合的目的是将不同来源的芯片数据进行处理，使得相同基因的数据可以相互比较。在预处理过程中，不同的标准化方法会影响不同来源的芯片数据之间的可比性。Autio 等比较了来自于 5个芯片组的 6926个基因表达数据，评估 5 种标准化方法的应用效果。经过研究发现，采用 AGC 方法(array generation based gene centering normalization)先进行样本内标准化再进行样本间的标准化时，能够得到最好的预处理结果，即在数千个样本之间得到可比较的基因表达量。此外，Stafford 等从以下 3方面对 8 种常用的标准化方法进行比较：敏感性和通用性、功能/ 生物学解释以及特征选择和分类错误，方便用户挑选合适的标准化方法进行跨实验室、跨平台的基因芯片表达数据的比较。采用一定的统计方法对不同来源的芯片数据进行整合，可以在进行更少实验的情况下更好地利用已有芯片数据，有助于发现多种癌症样本中共同的生物标志物以及某种癌症特异的生物标志物，其中，最简单的方法是 Z 打分归一化。较复杂的方法是提取不同数据集中表达数据的分布特征参数，根据这些特定的参数进行数据集匹配，包括：DistanceWeighted Discrimination、Combatting Batcheffects、disTran、Median Rank Score、Quantile Discretizing 和 Z 打分变换等。其中，经典方法的是 Daniel 等最早提出的荟萃分析(Meta-analysis)方法。利用 ONCOMINE 数据库，他们收集了 40个独立数据集(超过3700 个芯片实验)，提出了一种独立于单个数据集的统计量 Q-value，寻找多种来源数据集中显著差异表达的基因作为荟萃标志物(Meta-signature)。此后，多基因芯片融合方法得到了普遍关注，各种统计方法被用于发现通用标志物并与 Meta-analysis 方法进行比较。例如，Xu 等收集和整合了 26个公开发表的癌症数据集，包括21个主要的人类癌症类型的1 500个基因芯片数据，应用 TSPG(top-scoring pair of groups)分类器和重复随机采样策略，识别通用的癌症标志物。评估结果表明，采用一定的统计方法整合多种芯片数据能够识别出更加稳健的癌症标志物，相比单基因芯片得到的标志物，其将癌症类型与正常组织的区分效果更好。

(三) 蛋白质组学方法

通常,功能蛋白的表达异常和调节异常是疾病发生的分子标志,这些决定个体生物性状、代谢特征和病理状况的特殊功能蛋白可以作为潜在的药物靶标。尽管 90%的已知药靶为蛋白质,但由于数据和技术上的原因,蛋白质水平的药物靶标并不如基因、转录水平的研究广泛。近年来,随着更多蛋白质详细数据的获得,在蛋白质水平上进行药物靶标的开发和验证成为研究的热点。

1. 基于蛋白质的理化特性　在蛋白质的理化属性、序列特征和结构特征上,药靶分子和非药靶分子存在着显著的差异。Bakheet 等的工作具有一定的代表性。他们系统分析了148 个人类药靶蛋白质和3573 个非药靶蛋白质的特性,寻找两者的区别并预测新的潜在药物靶标。人类药物靶标蛋白可以归纳为8个主要属性: 高疏水性、长度较长、包含信号肽结构域、不含 PEST 结构域、具有超过2个N-糖基化的氨基酸、不超过一个O-糖基化的丝氨酸、低等电点和定位在膜上。以这些特征作为支持向量机的输入,可以在药靶和非药靶类之间达到96%的分类准确率,并识别出668 个具有类似靶标属性的蛋白质。基于蛋白质的理化特性进行药物靶标预测,有利于发现药物靶标的一般特征,方法直接、简单,但该方法受已知药靶的影响较大,在确认药靶的有效性时还需要引入更多的证据支持。

2. 基于蛋白质相互作用的网络特征　癌基因(oncogene)是人类或其他动物细胞(以及致癌病毒)固有的一类基因,又称转化基因,它们一旦活化便能促使人或动物的正常细胞发生癌变。通常,癌基因作为网络的 hub蛋白参与多种细胞进程,在信号通路中间成为信息交换的焦点。发现新的癌症相关基因是癌症研究的主要目标之一,也是发现潜在药靶的基础。人类基因组规模的蛋白质相互作用数据的快速积累为研究癌基因在细胞网络中的拓扑属性提供了条件。在蛋白质相互作用网络的基础上, Xu 等提取了节点的 5个网络特征,包括连接度、1N指数、2N 指数、与致病基因的平均距离以及正拓扑相关系数(positive topology coefficient),采用 KNN 方法比较疾病相关基因和对照基因在网络特征上的区别。研究结果证实: 疾病相关基因具有更高的连接度,更倾向与其他的致病基因发生相互作用,而且致病基因之间的平均距离明显低于非致病基因。Ostlund 等通过筛选与已知癌基因高度连接的基因,得到了一个由 1891 个基因组成的集合。通过交叉验证、分析功能注释偏好性和癌症组织中的表达差异进行方法验证,提供了一个较为可信的癌症相关的候选基因列表。该基因列表的规模是已知癌基因数目的2倍以上,对于生物标志物和药靶发现具有一定的提示作用。进一步, Li等通过整合多种数据源识别癌基因,包括网络特征、蛋白质的结构域组成和功能注释信息等。这些研究表明: 根据蛋白质在相互作用网络中的特征,能有效地提示大量的潜在药物靶标,并且方便与其他方法相结合。同时,蛋白质复合物的拓扑属性和模块性也可用于药靶筛选。不同于一般的二元蛋白质相互作用,复合物更接近于细胞内的真实状态。在复合物内部,多肽之间相互连接成为不同的核,其他蛋白质与核发生相互作用形成各种模块。蛋白质相互作用网络体现了蛋白质组的系统水平描述,对于建模复杂的生物系统具有非常重要的作用。有关蛋白质相互作用的知识可以使人们在分子水平上更好地理解信号转导的生理学活动,以及由于通路的交叠部分异常造成的多种疾病。

3. 比较蛋白质组方法　蛋白质组学是研究特定时空条件下细胞、组织等所含蛋白质表达谱的有效手段,也是寻找癌症分子标记和药物靶标的重要方法。相关的蛋白质组学技术包括免疫亲和纯化(affinity purification)、蛋白质活性表达谱(activity-based profiling)和蛋白

质芯片（microarray）等，识别与某一特定疾病或者病理条件相关的蛋白质。基于蛋白质组学研究药靶通常采用比较蛋白质组分析方法，例如，稳定放射性核素差异标记、ICAT（isotope-coded affinity tag）或 iTRAQ 技术，能够较为精确地定量蛋白质丰度的变化。通过比较癌症人群与正常人群在对应病理组织/器官内蛋白质的差别，挖掘潜在的药物靶标。例如，Hu 等采用二维液相色谱串联质谱法（2D-LC-MS/MS）比较肺癌患者与正常人的血清蛋白差异，经过蛋白质鉴定和定量分析，发现了2078 个蛋白质可能存在差异，进而挑选出Tenascin-XB（TNXB）作为候选的生物标志物用于预测肺癌的早期转移。此外，如果不能直接找到对应的活性小分子，也可以通过比较疾病样本和正常样本中蛋白质的表达差异，鉴别发生异常的生物学通路。采用总体的蛋白质谱方法（如MudPIT）获取充足的信息，发现与特定表型相关的蛋白质和通路。定位到相应的生物学通路之后，再从中确定药物靶标。随着人类蛋白质组计划的推进，蛋白质组技术的发展为系统地、规模化地寻找蛋白质药靶和蛋白质药物提供了有力的武器。但由于现有数据的规模和质量问题，以及分析方法的限制，采用蛋白质组学方法发现的药物靶标还没有人们预想的多，还有着广阔的发展空间。

(四) 代谢组方法

代谢组学是生物体内小分子代谢物的总和，所有对生物体的影响均可反映在代谢组水平。代谢组放大了蛋白质组的变化，更接近于组织的表型。代谢途径的异常变化反映了生命活动的异常，因此定量描述生物体内代谢物动态的多参数变化可揭示疾病的发病机制。通常，代谢组学的实验技术包括磁共振、质谱、色谱等，其中磁共振技术是最主要的分析工具，其次是液相色谱 - 质谱联用（LC/MS）和气相色谱 - 质谱联用（GC/MS）。通过GC/MS 技术解析出代谢物的质谱图，将其与现有数据库进行比较，可以鉴定该代谢化合物。由于缺少标准的代谢物数据库，该方法的鉴定结果有限。采用生物信息学方法对代谢组数据进行分析和处理，比较正常组和模型组的区别，可以帮助药靶发现以及药效评估。如 Pohjanen等提出了一种名为统计多变量代谢谱（staistical multivariate metaboliteprofiling）的策略，在代谢 GC/MS 数据的基础上辅助药靶模式发现和机制解释。同时，代谢组学对于生物标志物发现、药物作用模式和药物毒性研究具有重要作用。在酶网络的基础上，Sridhar 等发展了一种分支定界（branch-and-bound）方法，命名为 OPMET，寻找优化的酶组合（即药物靶标），用于抑制给定的目标化合物并减少副作用。类似的，通过提取代谢系统的特征，Li等采用整数线性规划模型在整个代谢网络范围内寻找能够阻止目标化合物合成的酶集合，并尽可能地消除对非目标化合物的影响。

(五) 整合多组学数据的系统生物学方法

系统生物学将基因组、蛋白质组和代谢组等不同组学的数据进行整合，研究在基因、mRNA、蛋白质、生物小分子水平上系统的生物学功能和作用机制。对于疾病的发生和发展提供了更好的理解，同时有助于识别药物的作用和毒性、模拟药物作用的过程、发现特异的药物作用靶标。

1. 文本挖掘方法　由于人类疾病背后的生物机制相当复杂，在药靶发现中最重要的任务不仅是要挑选和优化可靠的作用靶点，而且要理解在疾病表型下隐含的分子相互作用，提供可预测的模型并建立人类疾病的生物网络。因此，需要广泛地收集和过滤现有的各个

层面的异质数据和信息。目前，最流行的生物医学文献数据库 MEDLINE/PubMed 收录了从 1970 年开始的超过 1800万篇文献的摘要，并且每月还会新增超过 6 万篇的摘要。据估计，存储化学、基因组、蛋白质组和代谢组数据的数据库规模每两年就会翻一倍。如此丰富的生物数据和信息为药靶发现提供了巨大的新机遇。

尽管分子生物学和医学研究中数据库的重要性日益增长，绝大部分的科学论文并非存在于结构化的数据库条目中。这些知识必然无法为计算机程序所理解，甚至对于人来说都是难以发现的。文本挖掘方法是机器学习和自然语言处理方面的计算方法，能够有效地用于数据挖掘和知识理解，从海量的医学文献中挖掘与药靶发现相关的有用知识。

其主要内容包括：识别生物学实体，包括基因、基因产物、通路和疾病；提取蛋白质相互作用关系，并以网络图形化表示；抽提出特定细胞类型中相关的生物学通路，以及计算机仿真所需的动力学参数，建立存储这些抽提信息的数据库。目前，生物知识的文本挖掘方法主要采用实体的共出现分析和自然语言处理，已成功地用于疾病相关的网络重建以及生物数据分析，常用软件包括 Protein Corral和EBIMed。进一步，更复杂的文本挖掘方法可以从文献中抽提详细的相互作用注释信息，如Wang等发展了一种 CMW（correlated method-word）模型从文本中提取蛋白质相互作用的检测信息。

2. 通路建模与仿真　药物作用是一个复杂的动态过程，如果不能找到合适的方法就很难确认药物的有效性。例如，在药物开发过程中常用的手段之一是基因敲除实验，其作用方式与在特定酶上的竞争抑制过程完全不同。在基因敲除过程中，给定的通路可能被完全关闭，也可能由于系统的自身补偿作用而只有部分的影响。在此基础上设计的靶向药物可能存在效率较低的问题。因此，为了使药物开发过程更贴近于真实情况，有必要将定量的建模方法引入到药物研究领域，精确地模拟药物与靶标相互作用进而发挥药效的过程，发现更加有效的药物作用靶点。随着实验技术的发展、数据的累积和文本挖掘的开展，生物通路的建模方法得到了快速的发展和应用。其中，最常用的建模方法是确定性生化反应描述，已成功地用于药物代谢动力学和药剂反应建模。确定性反应的缺点在于缺乏可伸缩性。通常，基因组和蛋白质组方法要处理数十甚至数百个分子之间的信号网络，反应参数的范围可能包含多个跨度，超出了确定性方法的处理能力。最新出现的方法，如结合反应（combinatorial reaction generation）和线性规划（linear programming）可以满足这种需求，批量地处理大规模的复杂化学反应网络。进一步，随机方法能够从根本上克服确定性方法的限制。它们是高度可伸缩性的，同时易于进行模拟。然而，面对复杂的非线性动态问题，随机方法也存在很大的难度，还有待进一步探索。近年来，用于描述反应动力学网络的数学模型被证明可以有效地预测生物体对于环境刺激和外界扰动的响应，识别可能的药物靶标。一种系统的药物设计方法是：在网络中模拟单个反应的抑制过程，量化在指定观察量上的作用效果。在代谢网络中，观察量一般是稳态值，在信号级联模型中，观察量包括浓度、特征时间、信号持续时间和信号幅值等。Schulz 等在系统生物学标记语言Systems biology markup language（SBML）的基础上开发了一款名为 TIde 的工具，采用普通微分方程对系统进行模拟，研究在网络中不同位置进行激活和抑制处理时系统的响应。通过模拟不同的抑制目标、类型和抑制剂浓度，确定一个或多个优化的药物靶标，在尽可能少的抑制剂数目下以较低的浓度使指定的观察量达到期望值。此类药物作用模型的建立和模拟有助于理解药物的作用机制，预测药效发挥过程中可能存在的问题，进而为实验设计提供辅助作用。

3. 多组学数据的综合应用 系统生物学的优势在于“整合”，即综合利用基因组学、转录组学、蛋白质组学和代谢组学研究药物对系统的影响，提示可能的作用靶点。例如，Chu等根据大规模实验及相关数据库建立了整合的蛋白质相互作用数据集，采用非线性随机模型、最大似然参数估计和Akaike信息准则（Akaike information criteria，AIC）方法，通过基因芯片数据估计疾病状态和正常状态下的蛋白质相互作用网络差异，识别受到扰动的枢纽（Hub）蛋白节点，发现候选的药物靶标。除将转录组和蛋白质组数据结合之外，基因组与转录组、基因组与蛋白质组甚至更多组学数据的整合研究也在进行中。整合研究的关键是以生物网络为中心加深对整个系统的理解。疾病是一个非常复杂的生理和病理过程，涉及多基因、多通路、多途径的分子相互作用的过程，这种网络化的特点对于药靶筛选至关重要。系统生物学为药物开发过程提供了全新的视野，将蛋白质靶标置于其内在的生理环境中，在提供网络化的整体性视角的同时不会丧失关键的分子作用细节。鉴于生物网络具有一定的冗余性和多样性，包括一定的反馈回路和故障安全（fail-safe）机制。因此，筛选潜在药靶时要考虑到其在网络中的位置，优先挑选那些处于枢纽位置发挥重要作用的靶点，并且避免反馈回路对药效进行补偿。同时，疾病相关网络的内部高连接度表明，基于网络的诊疗方法应以整个通路作为靶标，而不是单个蛋白质。其最终的目标不仅是识别一组能够共同发挥作用的药物，而且发现一组靶标或模块的组合，它们在不同的治疗位置发挥作用并最后集中到一个特定的通路位点。尽管看起来这是一个几乎不可能实现的任务，但是在乳腺癌转移上的实验已经证明了基于通路知识进行多靶点联合治疗的有效性。

二、潜在药靶的生物信息学验证

在大量的潜在药靶被揭示之后，在此基础上可以寻找针对性的抑制小分子，进行后续的动物实验、临床测试等一系列药物开发过程。由于药物开发的难度较大、周期很长，在前期对候选药靶进行充分的筛选和验证显得非常必要。生物信息学方法在对候选药靶进行功能分析、预测其可药性并降低药物副作用方面也有重要的应用。

（一）蛋白质的可药性

随着超过上百个真核和原核生物的基因组被完整测序，人们有机会对基因进行大规模的分析和筛选，据估计整个人类基因组中约有 10%与疾病相关，从而导致约 3000 个潜在的药物靶标。同时，还有成千上万个来自于微生物和寄生生物的蛋白质，可以作为传染病治疗的药靶。目前，在所有的人类基因产物中仅有 2%（260~400）成功地发展为小分子药物的靶标。从大量的潜在靶标中挖掘能够被疾病修饰的可药部分是药物靶标验证的重要环节。根据基因组信息和蛋白质结构特征，人们开发了一系列生物信息学方法预测潜在靶标的可药性。评估蛋白质可药性的第一步是识别在蛋白质表面的所有可能的结合位点，进而寻找真实的配体可结合位点。其计算方法主要分为两类：基于几何的方法和基于能量的方法。几何基础上的方法利用了这样一个事实：天然的配体结合位点在蛋白质表面倾向于内部凹陷，例如 SURFNET、LIGSITE、SPROPOS、CAST、PASS 和 Flood-fill 方法。而能量基础上的方法将多种物理指标综合到 pocket 识别过程，试图计算其结合能，如 GRID、vdW-FFT、DrugSite和 Computaional solvent mapping。在排序过程中，这些方法都能够给予真实的配体结合位点

以较高的打分，证实了其有效性。第二步是评估结合位点能否高亲和性、特异地与小分子药物结合。定量评估给定位点可药性的计算工具较少，最直接的评估蛋白质可药性的方法是根据生物化学谱实际测量小分子击中目标的数目和类型，如MR谱图。

此外，由于大部分的蛋白质是通过与其他蛋白质相互作用发挥生物学功能，蛋白质相互作用在组织的各种细胞过程中发挥了基础和关键作用，被认为是一种富于挑战的同时又充满吸引力的小分子药物作用的新型靶标。类似于单个蛋白质的可药性，人们提出了多种方法预测蛋白质相互作用的可药性。2007年，Sugaya等从3个方面评估蛋白质相互作用的可药性：蛋白质相互作用中包含的结构域对、蛋白质与小分子药物的结合位点、GO功能注释的相似性打分。最近，Sugaya 等使用结构、药物和化学以及功能相关的69个特征作为支持向量机的输入，判断1295对已知结构的蛋白质相互作用的可药性，在标准的相互作用数据集中得到了81%的预测准确率，其中区分度最大的特征是相互作用蛋白质的数目和通路数目。

（二）药物的副作用

多组学数据的大量累积为药物研究提供了发展机遇，人们开发了多种方法用于发现潜在的药物靶标，但是最终找到合适的药物作用靶标并成功地进行临床应用并非易事。一般选择药物作用靶标要考虑两个方面的情况：首先是靶标的有效性，即靶标与疾病确实相关，通过调节靶标的生理活性能够有效地改善疾病症状。其次是靶标的副作用，如果对靶标的生理活性的调节不可避免地产生严重的副作用，那么将其选作药物作用靶标也是不合适的。

药靶和药物代谢酶多态性是造成药物疗效差异和毒副作用的主要原因之一。药物反应个体差异与个体的基因多态性特别是单核苷酸多态性（singlenucleotide polymorphism，SNP）密切相关。SNP 主要是指在基因组水平上由单个核苷酸的变异所引起的 DNA序列多态性。SNP 在人类基因组中广泛存在，平均每500~1000 个碱基对中就有 1 个，估计其总数可达300万个甚至更多。事先确定药物靶标的基因多态性，就可以估计药物适用的人群，进行个性化的医疗，增加疗效并降低毒副作用。目前，随着快速、规模化技术的发展，大量的 SNP 已经被揭示，为相关研究提供了基础。而生物信息学方法可以帮助阐释 SNP 与疾病治疗之间的关系，发现疾病易感基因和潜在药物靶标，评估药物疗效和毒副作用。以乳腺癌为模型，Wiechec 等报道SNP基因型会影响DNA 修复基因的转录活性和药物代谢过程，从而影响到临床的治疗毒性和效果。

在生物网络基础上综合评估药物作用的多种影响，也有助于寻找增加药物疗效、降低副作用的有效方法。在蛋白质-药物相互作用网络的基础上，Xie 等介绍了一种新的计算策略识别基因组规模的蛋白质-受体结合谱，用于阐释CETP抑制剂的药物作用机制。通过将药物靶标与生物学通路相关联，揭示了CETP抑制剂的副作用受多个交联通路的联合控制，给出了降低此类药物副作用的可能方法。

随着大规模组学数据的积累，仅凭实验方法已经不能满足数据分析和药靶发现的需求，有必要发展有效的生物信息学方法存储、分析、处理和整合多组学数据，提高药靶发现和验证的效率。目前，生物信息学方法已成功地运用于药靶发现的各个环节，对于存储疾病相关的医学数据、发现大量潜在的药物靶标、揭示药物作用机制、评估作用靶点的可药性等方面作出了重要贡献，有利于设计更加有针对性的生物学实验，促进现代新药开发进程。相比其

他方法,采用生物信息学预测潜在药物靶标的优势在于: ①不局限于特定的技术或某种类型的信息,尤其适合将不同的数据整合到一个大的体系中评估潜在药靶的表现; ②以网络为基础的药靶发现平台有利于从整体角度进行药靶筛选并发现联合靶标; ③随着动态的详细的生物学时空数据的累积,有可能在计算机中精确地模拟药物针对靶标作用的过程以及对整个系统产生的影响,从而大大提高药物开发的效率。

生物信息学方法在药物靶标发现的应用还刚刚起步,有赖于生物学理论、实验技术、统计分析和建模方法等多方面的进一步发展,从而在后基因组时代的疾病诊断、预后和个性化医疗中发挥更加重要的作用。

第三节

药物基因组学

Section 3 Pharmacogenomics

一、概述

所谓药物基因组学，就是在人类全基因组范围研究药物反应差异的遗传学本质。人类基因组具有广泛的多态性，药物基因组学研究个体的遗传背景，预测药物反应特点，实施“个体化”合理用药，并可以根据不同人群及不同个体的遗传特点设计、开发和研制新药。

药物基因组学以基因多态性为基础，而基因多态性是指群体中正常个体的基因在相同位置上存在差别（如单碱基差别、或单基因、多基因以及重复序列数目的差别等），并且这种差别出现的频率大于1%。药物基因组学研究药物效应的个体间差异，针对不同个体基因型进行个性化治疗。其研究内容包括药物效应的基因型预测和基因组学在医药上的应用，在分子水平上证明和阐述药物疗效、药物作用的靶位、作用模式和毒副作用。

药物基因组学不是以发现新的基因和探索疾病的发生机制为主要目标，而是以探讨药物作用的遗传分布，确定药物作用靶点来满足临床上最佳的药物效应及安全性为目标。药物基因组学除了研究遗传多样性引起的药物或有毒物质反应的差异外，还研究基因多样性与药效的关系，以及个体差异与同种药物不同作用靶点的关系等。

二、药物基因组学的生物标记

（一）单核苷酸多态

人群中大多数可观察到的序列突变是由单核苷酸多态导致的。单核苷酸多态（single nucleotide polymorphisms，SNP）是碱基对的变异，大约每1000bp的DNA序列中会有一个SNP。一个特定基因的SNP位置将会决定对这个基因功能可能造成的影响。

1. 基因编码区的SNP　大约有1%的SNP影响DNA序列中蛋白质编码部分，这些SNP通常叫做编码SNP（cSNP），它们的存在可能导致蛋白质多肽链中氨基酸序列的改变，对蛋白质的功能有影响，影响的范围可以从没有明显影响，增强活性，到蛋白质功能的完全失活。在许多例子当中，这种影响往往是不明显的，至少在基于表达的重组蛋白质功能分析上，常常通过增强或者阻碍蛋白质的降解过程来间接影响蛋白质功能。

SNP除了将一个氨基酸替换成另一个氨基酸之外，还能产生其他影响。比如，可能导致

一个DNA遗传密码子过早地终止蛋白质翻译，这个被称之为一个过早的终止密码子或无意义密码子。这种密码子会导致一个截断的蛋白质迅速降解或功能失活。有趣的是，最近的证据表明，这样的截断蛋白质可能会一直形成，不会结束，这是因为有一个系统监控mRNA中无意义密码子的存在，导致mRNA在降解之前翻译。*CYP2C19*3*等位基因就是一个典型的过早终止密码子的例子。*CYP2C19*基因编码的细胞色素P450 CYP2C19蛋白，是一种肝药酶，主要负责代谢多种药物，包括巴比妥类、乙内酰脲和多种质子泵抑制剂（如奥美拉唑）。一部分个体体内的这种酶含量不足（代谢不良），将影响给药和疗效。比如说，用奥美拉唑和阿莫西林治疗与幽门螺杆菌感染相关的胃溃疡和十二指肠溃疡的患者，不良代谢者比快速代谢者的治疗效果更好。然而，代谢不良的发生因人群而异，相较于欧洲白种人后裔（2%~5%），东亚人发生的频率更高（18%到23%，包括日本，中国和韩国）。这些人群的差异部分源于*CYP2C19* 3*等位基因，其中发生在东亚人群中的约6%至10%，但对于欧洲白人后裔基本上不发生。与*CYP2C19*3*等位基因相关的SNP导致色氨酸密码子TGG被替换成终止密码子TGA，这就导致了酶被截断，从而不稳定以及失活。

在大多数情况下，在蛋白质编码区的SNP对氨基酸编码没有影响。这些类型的SNP被称为同义（或沉默）SNP，它们是退化氨基酸编码系统的结果，这个系统被蛋白质翻译机器所使用，使得DNA核苷酸三联体可以有几种不同的组合编码同一种氨基酸。虽然同义SNP通常被认为对蛋白质功能没有影响，但是最近一项研究却给出了相反的结果。人类MDR1基因（编码P–糖蛋白）中发现了在外显子26中有一个常见的同义SNP，这个SNP先前与活体内P–糖蛋白活性的降低有关，当各种哺乳动物细胞系中的重组蛋白质被表达时，它就改变P–糖蛋白的构成和底物的特异性。这个结果是由于在参照序列中比较常用的甘氨酸密码子GGC被突变序列中相对较少的甘氨酸密码子GGT所替换，这样的翻译效率可能比较低，导致翻译的蛋白折叠区时间的改变，以及对蛋白质的功能有不利的影响。

SNP的位置位于一个基因的编码区域之外也会造成一定的影响，包括对基因转录、mRNA的剪切、mRNA的降解的影响，以及蛋白质翻译效率的影响。

2. 启动子和5’–调控SNP　位于一个基因启动子和5′–调控区域的SNP可以影响（增加或减少）基因的转录，这是通过DNA序列中的变化实现的，结合转录因子很有必要，这些转录因子对底物的活性、基因表达的增强与抑制都很重要。作为后者的一个例子，位于一个假定γ–干扰素激活序列元件的SNP用来消除γ–干扰素对CYP2E1转录的正常抑制影响，γ–干扰素激活序列位于CYP2E1基因的5′–调控区域。CYP2E1是一种药物代谢酶，它催化各种低分子量药物的氧化，最显著的是乙酰氨基酚（高浓度），导致肝毒性代谢产物N –乙酰– P –苯醌亚胺的形成。

3. 内含子和剪切SNP　转录开始之后，主要的mRNA转录通常包含编码区（外显子）和非编码区（内含子）的mRNA。内含子中的mRNA通常通过酶切被移到一个更短的成熟mRNA中来作为蛋白质翻译的模板。一些保守序列元件已经被证明对于mRNA在特定位点的剪切是必不可少的，包括在外显子和内含子边界处供体和受体位点的剪切，以及分支位点的剪切。也有一些位于外显子和内含子区域中保守度较低的序列元件，处在剪切位点的周围，这些剪切位点也为调控蛋白提供结合位点，调控蛋白既可以促进（剪切增强子）或者抑制（剪切抑制子）mRNA的剪切。后者的调控位点提供了一个机制，通过这个机制，不同的mRNA剪切产物可以从相同的主要mRNA转录的调控方式中产生。

位于上述mRNA序列元件的任何一个中的SNP都有可能改变mRNA的剪切，导致mRNA缺少一个或更多外显子，或者部分地或完全地保留内含子序列。另外，SNP可能创造新的剪切供体、受体、或调控位点，这些调控位点同其他剪切位点竞争，然后再形成新的mRNA突变。虽然产生的剪切形式可能编码一个功能蛋白突变，但是更可能的情况是，这样的变化导致过早终止密码子和一个没有活性的蛋白质的出现。

与*CYP3A5*基因有关的SNP说明了上述几个可能性。CYP3A5是临床上重要的酶CYP3A亚科（其中还包括CYP3A4，CYP3A7和CYP3A43）的其中一种，它负责人体中大部分药物的氧化。*CYP3A5*在人体中有多种表达状态，只有33%的欧美人和60%的美国黑人的肝脏里有这种酶的表达。与低*CYP3A5*表达关联的最常见的突变等位基因是*CYP3A5 * 3*。*CYP3A5 * 3*包含了一个SNP，这个SNP在*CYP3A5*基因的3号内含子中创造了一个新的剪切位点，导致异常的mRNA转录发生在了3号内含子（外显子3B）的部分区域。3B外显子包含了一个出现在102位氨基酸之后的过早终止密码子，导致了一个截断和失活的*CYP3A5*的出现。与此相反，有些罕见的主要出现在美国黑人中的*CYP3A5*6*等位基因与一个SNP（g.14690g>a）有关，这个SNP破坏了7号外显子内剪切增强子的结构，导致了7号外显子彻底地从成熟的CYP3A5的mRNA中删除。由此产生的转录在183位氨基酸之后出现一个阅读框和过早终止密码子的转换，随后就导致一个截断的、失活的CYP3A5蛋白质出现。

4. mRNA UTR SNP　最后，SNP可能位于成熟mRNA的5’端或3’端未翻译的调控区（UTR）。虽然未翻译的mRNA不能编码蛋白质，但是对于许多基因，这个区域可以通过二级结构的形成与调控蛋白产生互作，增强或阻碍mRNA的降解来改变mRNA的稳定性或直接影响蛋白质的翻译效率，这些都影响了蛋白质的形成速率。因此，破坏RNA二级结构的SNP可对这样的RNA和蛋白互作产生不利的影响。

一个已经被证明的例子是，*TYMS*基因中的D等位基因通过改变RNA和蛋白质互作来影响基因的表达。该基因编码胸苷酸合成酶，它是胸苷酸从头合成的关键酶，也是氟尿嘧啶等抗癌药的治疗靶点。D等位基因不是严格意义上的SNP，它是一段由6个核苷酸构成的DNA伸展区（编码RNA），这段DNA被插入到TYMS基因中的3’-UTR区域，而这种情况在大约30%~40%的欧洲白人中存在。各种研究表明，结合并促使mRNA衰变的核蛋白AUF-1对TYMS的D等位基因编码的mRNA有着高度的亲和力，从而导致在TYMS的mRNA表达量降低。

此外，近年来的研究提示一种新的值得关注的机制：位于mRNA非翻译区的SNP通过改变miRNA的结合位点从而影响基因表达。miRNA是内源性基因的产物，有着非常短的RNA分子，这些RNA分子通过RNA干扰机制调控基因的表达，加速RNA的降解，抑制其翻译，或者两者都有。一个典型的例子是，乳腺癌细胞中高表达的*miRNA-328*通过抑制转运子ABCG2增强了米托蒽醌的敏感性。

（二）插入缺失和微卫星多态

尽管不如SNP普遍，但是其他类型的基因突变对药物基因组也有着重要的影响。其中最常见的包括简单DNA序列的插入和缺失，其范围从一个单个的DNA碱基到整个基因。小的缺失往往通过和SNP一样的方式影响基因的功能，相比较于简单的替换，碱基被增加或是移除可能会有更大的可能性影响基因的功能。

在基因的蛋白质编码区内，即使是相对较小的缺失也会对蛋白质的功能产生严重的影

响，尤其是当插入或删除的核苷酸数目不是3的倍数时。在后者的情况中，就会出现阅读框转变（移码突变），使得在翻译的过程中，造成碱基缺失之后，所有的氨基酸序列都被很大程度上改变，而且总是遇到过早终止密码子。

缺失的一种特定类型就是微卫星多态性。微卫星（核苷酸重复）是一段DNA区域，其中包含典型的重复序列，这种重复序列包括单核苷酸重复（tttttttt），二核苷酸重复（tatatatatata），三核苷酸重复（tactactactac）以及更高形式的重复。这些重复区域因为DNA的多态性，都是常见位点，这导致了重复的扩张，可能与它经过的DNA聚合酶产生的错误有关。

最被人所熟知的二核苷酸重复多态性与药物基因组学相关的例子就是Gilbert综合征，在大约10%的欧洲血统的白人中可以发现轻度未结合的高胆红素血症。这种综合征重复扩张的tata盒有关，tata盒位于编码UDP-葡萄糖醛酸转移酶（UGT）1A1基因的启动子内，这种酶主要负责胆红素的结合。在大多数人体内，在5´末端的tata盒序列延长至包含了6个"ta"的重复，然而在Gilbert综合征的患者中有7个（有时8个）"ta"的重复，这导致了UGT1A1基因转录较少。有趣的是，也有些人只有5个"ta"重复，相较于有6个"ta"重复的人，前者的转录水平更高。

虽然通常由于很低的UGT1A1水平导致的轻度未结合胆红素血症几乎没有临床症状，但是UGT1A1也是很重要的，它涉及一些重要药物的代谢，有些药物有相对较低的治疗指数。例如，SN-38经过UGT1A1的葡萄糖醛酸化以后就失去活性，SN-38是伊立替康的活性代谢物，伊立替康是一种治疗转移性结肠癌的药物。一些临床研究表明，携带UGT1A1*28等位基因的患者有更大的风险患有骨髓抑制、痢疾及其他不良反应，这可能是由于缓慢消除和增加SN-38作用的结果。然而，最近的研究表明，*UGT1A1*28*基因型的毒性预测值可能会减少重复的伊立替康给药。在*UGT1A1*28*纯合子患者中观察到的肿瘤应答率（药物疗效的指标）会更高，可能是由于癌组织中增强SN-38的作用。由此推断，*UGT1A1*28*可能影响药物的疗效和毒性。

（三）基因拷贝数多态性

相对于前面提到的两类多态，更大规模的DNA 重排，会导致部分或整个基因的缺失或复制。整个基因的缺失和复制属于基因拷贝数突变。

CYP2D6基因是这种重要的基因结构改变的典型例子，这种改变对许多临床上很重要的药物的代谢有重要影响。*CYP2D6*5*等位基因出现在1%~7%的人口中，它与整个*CYP2D6*基因的缺失，*CYP2D6*完全丧失活性以及不良代谢表型有关。相比之下，多达13个*CYP2D6*基因的功能拷贝已经被证实存在于一些个体当中，通过这种酶的代谢导致药物的超高清除率。虽然在亚洲、非洲黑人、欧洲白人的人群中相对罕见，*CYP2D6*基因的复制在某些东部非洲人群（埃塞俄比亚人）中高达29%。其他药物基因组学中重要的基因拷贝数多态的例子包括*CYP2A6*（在*CYP2A6*4*在15%的亚洲人口中缺失）、*UGT2B17*（在11%至12%的白人与黑人中缺失）、磺基转移酶1A1（在26%的欧洲白人和63%的非裔美国人中复制）、谷胱甘肽、S -转移酶M1（在50%至60%的白种欧洲人中缺失），谷胱甘肽- S -转移酶T1（在10%至15%的欧洲白人中缺失）。目前为止还未确定的基因拷贝数变异极有可能有助于药物基因组学发展。

（四）其他DNA序列突变

其他与药物基因组相关的涉及超过一个核苷酸差异的基因突变包括替换（一个序列与

另一个序列替换），倒位（反向副本替换原有序列），易位（从另一部分相同的基因或从另一个基因中插入序列）以及转换（一个基因中的部分序列变为另一个基因中的序列）。

*CYP3A7*1C*等位基因是一个转换的例子。这种突变可能是由于*CYP3A7*基因的一部分重要的启动子变为了*CYP3A4*基因的启动子。这种变化导致了DNA序列元件的增加，这种元件对于结合核受体转录因子孕烷-X受体（PXR）十分必要，在*CYP3A7*中通常不会发现雄烷受体（CAR）。尽管*CYP3A7*通常被认为是胎儿特定的亚型，但是携带*CYP3A7*1C*的个体（10%白种欧洲人）更可能在他们的肝脏和小肠中表达大量的CYP3A7蛋白，导致一些CYP3A底物更高的清除率。

（五）等位基因命名

DNA序列变异十分复杂，所以有必要具有一个相对简单而精确的系统来为基因组的序列突变命名。人类基因组变异协会（http: //www.hgvs.org/mutnomen/）已经提出这样的系统。在这个系统中，DNA序列中核苷酸变化的描述是相对于序列数据库中参考序列而言的。例如，一个常见SNP"CYP2B6 g.15631g>t"，*CYP2B6*是一个包含变异的基因，"g"表明它是基因组上的DNA，"15631"表明变异碱基对出现的相对位置，通常从基因中编码蛋白起始密码子的第一个核苷酸算起。"g>t"表明在参照序列中，鸟嘌呤G被变异序列中的胸腺嘧啶T替换。一般地，DNA参照序列应来自于RefSeq数据库，而且应该包含数据库登录号、版本号还有序列中起始密码子腺嘌呤的位置。这个系统稍作变化就可以用来描述基因产物的变化，比如"CYP2B6 r.516g>u"（g= 鸟嘌呤，u=尿嘧啶），"CYP2B6 c.516g>t，"以及"CYP2B6 p.Q172H"（Q=谷氨酸，H=组氨酸），描述对于相同的CYP2B6 SNP，在预测的mRNA、cDNA以及蛋白序列中相应的变化。

目前，dbSNP数据库收录了主要的人类SNP以及短的插入或缺失变异。在该数据库中，为了识别和检索，所有的已核实的SNP已经被指定了一个参照SNP号码以rs开头。例如，CYP2B6 g.15631g>t的RefSNP ID就是rs3745274。这些SNP数据，包括人类基因组序列还被整合到其他数据库中，以至于能够对于一个给定的基因，我们可以识别它的绝大多数SNP。dbSNP还提供了其他相关信息，包括不同种族和群体的SNP类型（编码还是同义），一些SNP的基因型，和等位基因频率。

这个系统恰当地描述了基因序列的单碱基变异，但是当描述同一个等位基因的多个序列变异（如单体型）时，便显得过于繁复了。因此，人类基因组组织（HUGO）已经提出了下面的等位基因命名系统，它能够描述复杂的等位基因突变。这个系统已经被CYP等位基因命名委员会采纳，命名新等位基因的完整规则可在网上获得。简单地说，"参照"基因序列被指定为1，比如*CYP2B6*1*，这就是*CYP2B6*基因的参照序列。相应的蛋白质产物被称为CYP2B6.1。如果第一个序列的变异确定会有功能相关的影响，它将被指定为*2，随后的突变被指定为一个新数字，这个数字按提交的先后顺序依次增加，比如在*CYP2B6*2*、*CYP2B6*3*、*CYP2B6*4*、*CYP2B6*5* 和*CYP2B6*6*中的 *3、*4、*5、*6等。相应的蛋白质变异为CYP2B6.2，CYP2B6.3，CYP2B6.4，CYP2B6.5和CYP2B6.6。对于其他尚无法确定功能的基因变异，例如与已知SNP连锁的同义SNP、内含子中的变异、启动子中的变异等子群，可以通过添加一个字母的方式来区分标注，比如CYP2B6*6B。

目前，基因或特定疾病的命名委员会已经负责命名新的等位基因变异，也负责维护已

知等位基因的可获得参考文献列表。大量特定基因变异数据库的列表可以通过URL http://www.genomic.unimelb.edu.au/mdi/dblist/glsdb.html获得。

三、药物基因组学生物标记的预测

有两种主要的方法可用作预测影响药效的基因组变异(即SNP等生物标记)——候选基因分析和全基因组分析(图12-10)。这两种分析方法中，DNA样本是从给药人群中获得的，

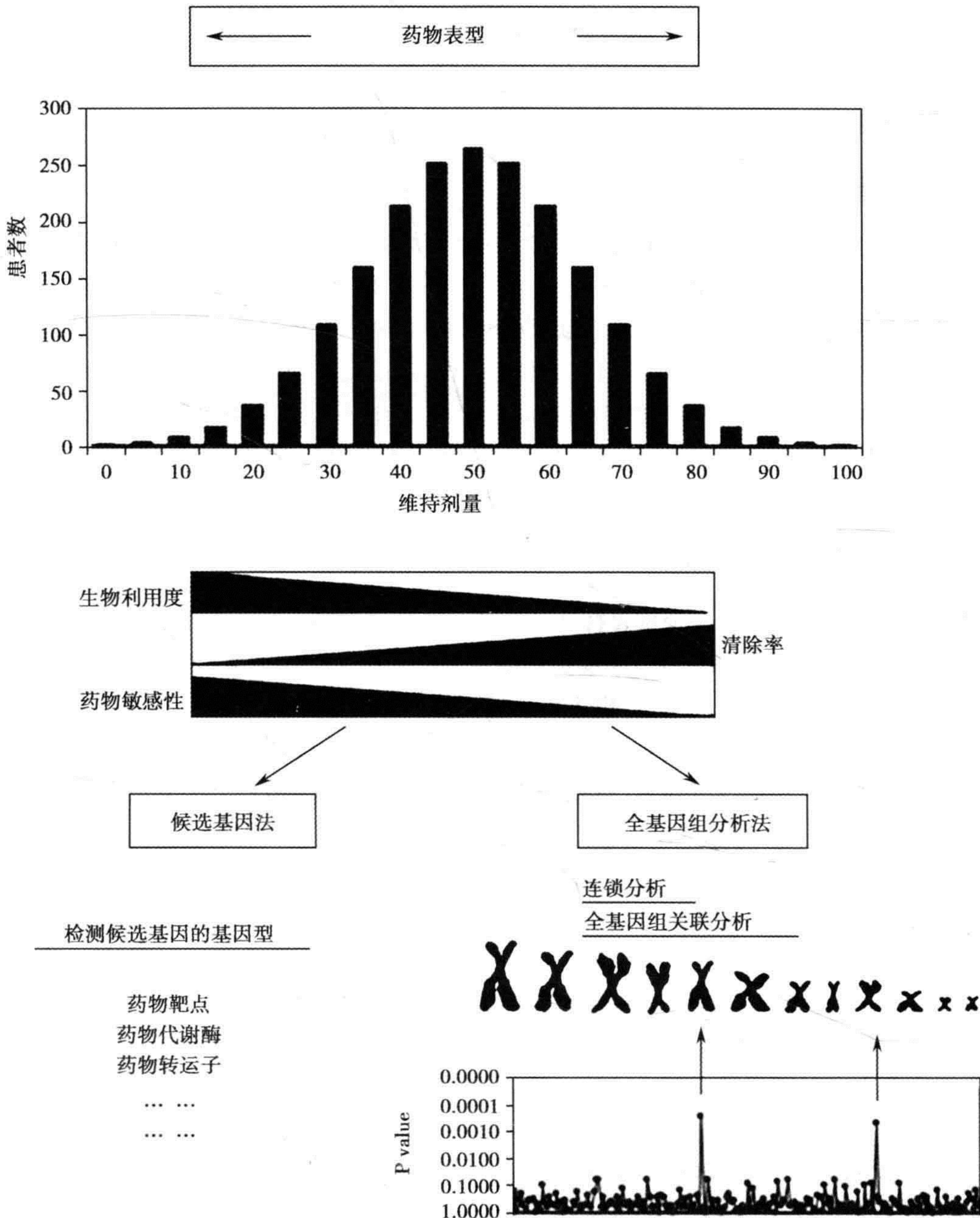


图12-10 识别药物基因组学生物标记的方法

这些人需要在一个事先确定的时间点检测药物的反应状态（有效、无效、不良反应等）。从中得出的结果变量称为药物表型（drug phenotype）。表型可以是二元变量（如有效和无效），也可以是连续变量（如患者维持药物剂量）。

（一）候选基因分析法

这种方法有一个前提条件，就是需要事先知道参与药物反应、影响药物效应的基因。一般地，候选基因包括编码药物代谢酶的基因、编码吸收和排除的转运蛋白的基因、编码药物靶标受体或酶的基因。该方法通常是先确定给药个体的表型，然后检测候选基因中SNP等遗传变异在给药群体中的状态，最后应用统计学方法分析表型—基因型间的关系，从而鉴别可用于预测不同药物反应的遗传学标记。通常，这些分析也能够解释种族、年龄和性别等其他重要的协变量。

（二）全基因组分析方法

全基因组的连锁不平衡分析是遗传学中鉴别致病基因的经典方法，目前也提倡用这类方法发现新的药物基因组学生物标记。这类方法不需要药物应答机制的先验知识，因此，它们会发现一些之前与感兴趣药物没有联系的基因。最初的分析对象是一个具有良好特性的遗传标记集合，通常是分布在基因组中的多态性位点（微卫星和SNPs），其基因型结果跟药物表型是相关的。一个与表型高度相关的基因组区域会被检测出具有高密度的遗传标记，直到一个特别地基因被检测出来，从这个基因序列中可能会直接的识别出新的遗传变异，这些遗传变异是和药物表型预测是高度相关的。

（三）连锁不平衡、单体型和标签SNPs

鉴别基因变异位点的一个重要特征是连锁不平衡。连锁不平衡是指在染色体上距离相对很近的位点倾向于共同遗传。高度连锁的基因变异位点称为单体型，在很多例子中单体型被排列在DNA不连续的区域中，称为单体型区块（haplotype blocks），这些区块是被基因重组的热点（gene recombination hotspots）分隔开的（图12-11）。因此，没有必要定位精确的、与表型成因有显著关系的变异位点，这些显著关系在同个单体型区块中的其他变异位点中也能找到。

随着高通量全基因组SNP微阵列的出现，使得一步完成全基因组基因型检测成为可能。虽然最终有可能实现对基因组中全部SNP的鉴别，但是这个并不是必要的，因为如前面所说的，许多的SNPs是位于同一个单一型中的，因此对每个SNP进行检测是冗余的。一个非常严格（parsimonious approach）的方法就是对特别的单体型选出一个SNPs的最小集合，称为单体型标签SNPs（tag SNPs或 tSNPs）。图12-12展示了一种鉴别单体型中tSNPs的方法。为了绘制出全人类基因组的全部单体型，国际人类基因组单体型图计划（international HapMap project）已完成对tSNPs的鉴别。为了验证这种方法，一个由904个tSNPs构成的集合最近被鉴别出来，还发现其代表了参与人类药物代谢和处置的55个基因的大部分遗传变异，这意味着tSNPs将在识别新的药物基因组标记中有广泛的应用。

一个很重要的问题是从一个给出的参考人群中鉴别出的tSNPs在不同的人群中是否具

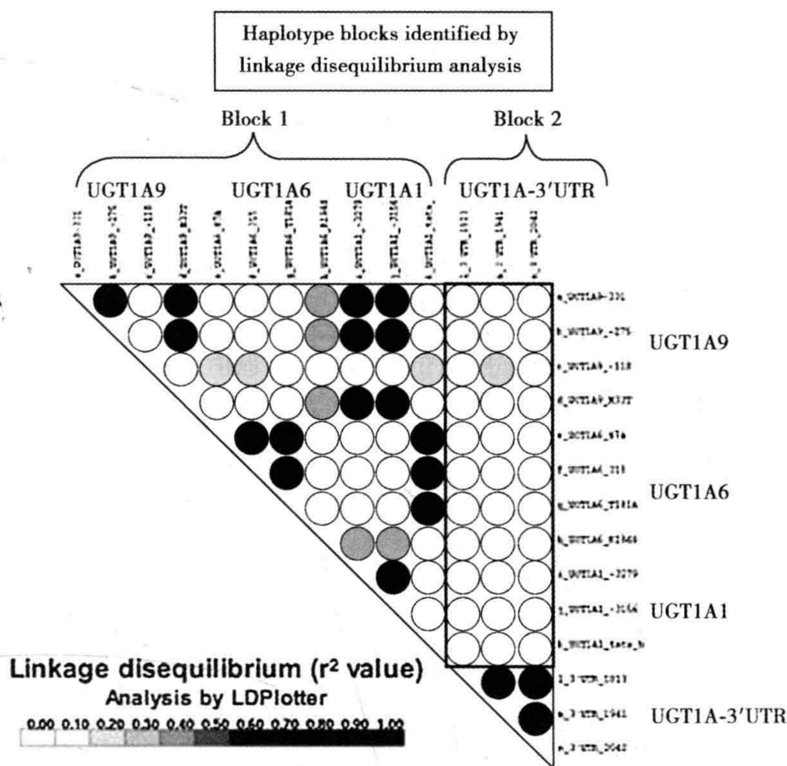


图12-11 连锁不平衡定义单体型块边界

本例中检测了54个白种人个体的*UGT1A*基因的SNPs，测序范围长约150 000bp，包括外显子1区，调控区5'-UGTs 1A9、1A6、和1A1，以及3'-UTR区。UGT的连锁不平衡应用软件LDPlotter(https://www.pharmgat .org/Tools)中的回归分析评价。图中，圆圈的灰度代表每对SNP比较的回归值。分析结果表明，两个不同的连锁不平衡区域，块1（UGT1A9, 1A6, and 1A1 SNPs）和块2（3'-UTR SNPs），具有块内高r^2和块间低r^2的特点。可见，分开的单体型块可以用于描述这些区域的遗传变异

有可转移性。一项关于52个不同人群的研究表明这种可能是存在的，从这些人群中鉴别出的83%的共有的单体型在产生国际人类基因组单体型图的人群中也是共有的。然而，一个重要的警告是随着假定的从非洲起源的人类祖先的遗传距离的缩小，单体型的多样性在增加。结果，很有必要补充一大部分共有的tSNPs和人群特异的tSNPs，特别是在那些很多近代非洲祖先的人群中。

在基因组范围内对表型与基因型联系的研究有一个重大隐患，是由多重检验引起的潜在的大量假阳性联系。例如，对500 000个SNPs进行单变量分析，如果采用α为0.05，理论上会有25 000个假阳性的结果。虽然假阳性率可以通过矫正p值来减小，如Bonferroni、FDR等多重检验校正，但是这同样也会消除一些比较弱的关联关系，而这些关系可能仍与表型的决定有关。折中的办法是，选择那些有最强关联的变异，同时通过文献或额外的实验来验证所发现表型相关基因的生物学意义。而且，在候选基因和全基因组方法中，非常重要的是要通过对不同患者人群的独立研究来证实找到的药物基因组标记的通用性。

由于全基因组分析方法主要是用相对常见的SNPs（通常是>5%的等位基因频率），这将

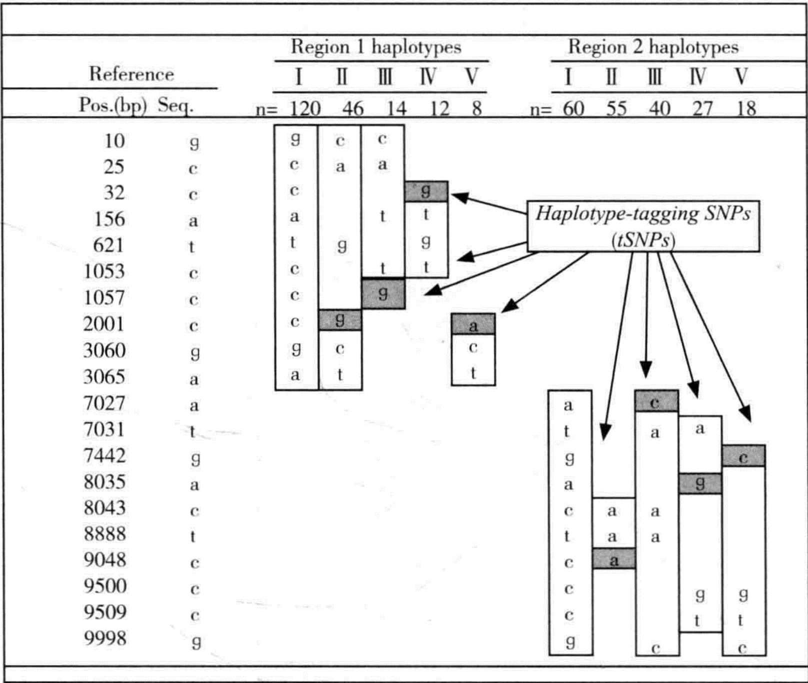


图12-12 单体型分析与tagSNP识别

本例中，一个约10 000bp长的基因在100个个体中测序，并应用程序fastPhase（http://www.stat.washingtonedu/stephens/software.html）推断单体型，结果识别出两个高连锁区域（区域1，10~3065bp；区域2，7027~9998）。每个区域中最常见得单体型与Genebank参考序列一致，定义为单体型Ⅰ。对于变异单体型Ⅱ到Ⅴ，不同于参考序列的SNPs特异地属于每个变异单体型（灰色）。*n*为被研究群体中具有该单体型的染色体数目

无法识别出罕见SNPs或其他序列变异。单个的罕见变异有潜在的能力（是从完全的基因失活来说）在一个特别的个体中对药物表型产生显著的影响，然而，多个罕见变异产生的累积效应有助于人群总体的表型多样化。因此，目前急需兼顾候选基因分析和全基因组分析两种策略的新方法。

（四）药物基因组学生物标记资源

目前生成的大量药物基因组数据，非常需要一个统一的资源来访问这些信息。许多在线的数据库提供预测信息来判定遗传变异与药物反应之间的联系。第一个也是目前最全面的一个数据库是药物基因组学和遗传药理学数据库PharmGKB（http://www.pharmgkb.org）。PharmGKB提供药物基因（pharmacogenes，如对药物反应有明确影响的基因）的超链接注释信息。这些信息是从多个资源中收集来的，包括发表的文献、NIH资助的遗传药理学实验，和其他一些学术和商业的资源。所有NIH资助的项目都要将其药物基因组的信息放到这个数据库中。

四、药物基因组学生物标记的应用

药物基因组学生物标记通过各种途径指导临床医师给药。这些生物标记的一个主要

用途就是判断一个患者的肿瘤是否会对一种特定药物产生反应，这是要看这个基因的表达产物是否会被该药物所靶向。例如，将病患体内EGFR过表达的检测结果，作为是否采用西妥昔单抗治疗转移性直肠癌和头颈癌患者的依据。另一个例子是，用曲妥单抗治疗乳腺癌患者前需要确认HER2/neu的过表达状态。FDA颁布的相关药物标签信息(drug label information)中明确指出，在给患者使用曲妥单抗等药物前必须检测相关的药物基因组生物标记。制定这样的规范是因为，最初认定的药物治疗作用数据是从被检测出阳性反应的患者中得到的，而后续研究的结果支持它们的有效性，例如多个研究报道了曲妥单抗对HER2/neu过表达阳性的乳腺癌患者具有更好的疗效。

最近，FDA建议，在药物使用前，应对两种新的药物基因组生物标记进行检测(但不是必须的)。它们是，*TPMT*(硫嘌呤甲基转移酶)的*2和*3等位基因和*UGT1A1*28*等位基因。携带*TPMT*纯合子变异的白血病患者在使用咪唑硫嘌呤或6-巯基嘌呤治疗时将增加严重骨髓抑制的风险，而携带*UGT1A1*纯合子变异的转移性结肠癌患者在使用伊立替康治疗是同样会增加骨髓抑制的危险。在这两个例子中，与杂合子和野生型患者相比，纯合子变异的患者的初始用药剂量显著减少，不过减少的程度没有明确的列出。在药物标签信息中提供的大多数其他药物基因组生物标记只是出于提供信息的目的，而不是关于检测目的。关于被批准药物标签的一个全面的、已证实的药物基因组生物标记的列表通过FDA的网站查询。

除了在药物治疗之前做检测之外，也有很多情况是在治疗过程中检测的。例如，在对HIV感染患者的治疗过程中，可以监测到病毒的基因型改变，并且在抵抗表型出现的时候调整药物用量。同样的方法也可以用于其他感染性疾病以及在癌症治疗中的耐药等情形。

(陈秀杰 徐建凯 刘 磊)

参考文献

1. Donna Maglott, * Jim Ostell, Kim D. Pruitt, et al. Entrez Gene: *gene-centered information at NCBI*. Nucleic Acids Res, 2007, 35(Database issue): D26-D31.
2. Zhenting Gao, Honglin Li, Hailei Zhang, et al. PDTD: *a web-accessible protein database for drug target identification*. BMC Bioinformatics, 2008, 9 : 104.
3. Daniel R Rhodes, Shanker Kalyana-Sundaram, Vasudeva Mahavisno, et al. Oncomine3. 0 : *genes, pathways, and networks in a collection of 18 000 cancer gene expression profiles*. Neoplasia, 2007, 9(2): 166-180.
4. Le Novère N, Bornstein B, Broicher A, et al. BioModels Database: *a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems*. Nucleic Acids Research, 2006, 34 (Database issue): D689-D691.
5. Brett G. Olivier, Jacky L. Snoep. *Web-based kinetic modelling using JWS Online*. Bioinformatics, 2004, 20(13): 2143-2144.
6. David J. Payne, Michael N. Gwynn, David J, et al. *Genomic approaches to drug discovery*. Curr Opin Chem Biol, 2006, 10(4): 303-308.
7. Nuria Conde-Pueyo, Andreea Munteanu, Ricard V Solé , et al. *Human synthetic lethal inference as potential anti-cancer target gene detection*. BMC Syst Biol, 2009, 3 : 116.
8. Hu G, Agarwal P. *Human disease-drug network based on genomic expression profiles*. PLoS One, 2009, 4(8): e6536.
9. Aravind Subramanian, Heidi Kuehn, Joshua Gould, et al. GSEA-P: *a desktop application for Gene Set*

Enrichment Analysis. Bioinformatics, 2007, 23(23): 3251–3253.

10. Reija Autio, Sami Kilpinen, Matti Saarela, et al. *Comparison of Affymetrix data normalization methods using6 926 experiments across five array generations*. BMC Bioinformatics, 2009, 10(Suppl 1): S24.
11. Phillip Stafford, Marcel Brun. *Three methods for optimization of cross-laboratory and cross-platform microarray expression data*. Nucleic Acids Research, 2007, 35(10): e72.
12. Daniel R. Rhodes, Jianjun Yu, K. Shanker, et al. *Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression*. Proc Natl Acad Sci USA, 2004, 101(25): 9309–9314.
13. Lei Xu, Donald Geman, Raimond L Winslow. *Large-scale integration of cancer microarray data identifies a robust common cancer signature*. BMC Bioinformatics, 2007, 8 : 275.
14. Bakheet TM, Doig AJ. *Properties and identification of human protein drug targets*. Bioinformatics, 2009, 25 (4): 451–457.
15. Xu J, Li Y. *Discovering disease-genes by topological features in human protein-protein interaction network*. Bioinformatics, 2006, 22(22): 2800–2805.
16. Gabriel Östlund, Mats Lindskog, Erik L. L. Sonnhammer. *Network-based identification of novel cancer genes*. Mol Cell Proteomics, 2010, 9(4): 648–655.
17. Li Li, Kangyu Zhang, James Lee, et al. *Discovering cancer genes by integrating network and functional properties*. BMC Medical Genomics, 2009, 2 : 61.
18. Xiaofang Hu, Yan Zhang, Aili Zhang, et al. *Comparative serum proteome analysis of human lymph node negative/positive invasive ductal carcinoma of the breast and benign breast disease controls via label-free semiquantitative shotgun technology*. OMICS, 2009, 13(4): 291–300.
19. Sleno L, Emili A. *Proteomic methods for drug target discovery*. Curr Opin Chem Biol, 2008, 12(1): 46–54.
20. Elin Pohjanen, Elin Thysell, Johan Lindberg, et al. *Statistical multivariate metabolite profiling for aiding biomarker pattern detection and mechanistic interpretations in GC/MS based metabolomics*. Metabolomics, 2006, 2(4): 257–268.
21. P Sridhar, B Song, T Kahveci, et al. *Mining metabolic networks for optimal drug targets*. Pac Symp Biocomput, 2008, 13 : 281–302.
22. Z Li, RS Wang, XS Zhang, et al. *Detecting drug targets with minimum side effects in metabolic networks*. IET Syst Biol, 2009, 3(6): 523–533.
23. Yongliang Yang, S. James Adelstein, Amin I. Kassis. *Target discovery from data mining approaches*. Drug Discov Today, 2009, 14(3–4): 147–154.
24. Wang Hongning, Huang Minlie, Zhu Xiaoyan. *Extract interaction detection methods from the biological literature*. BMC Bioinformatics, 2009, 10(1): S55 .

中英文名词对照索引

3' 非翻译区		285
5' 非翻译区		285
β链	β-strand	154
β球蛋白	beta-globin	44
β-折叠	β-sheet	154
β-转角	β-turn	155
Ω环形	Ω loop	155
AM	association method	198
BLAST	basic local alignment search tool,基本局部联配搜索工具	40
BLOSUM	block substitution matrix	36
cDNA芯片	cDNA microarray	76
DNA甲基转移酶	DNA methyltransferase	457
DNA甲基转移酶3 Like	DNMT3L	466
DNA甲基转移酶3A	DNMT3A	466
Donna Maglott		523
FN	false negative	112
FP	false positive	112
general protein/mass analysis for windows	GPAMW	164
GPCR mode	G蛋白偶联受体模式	181
*k*近邻	*k*-nearest neighbor	108
*k*均值聚类	*k*-means clustering	100
MIM	Mendelian inheritance in man	369
MLE	maximum likelihood estimation	198
*n*倍交叉证实	*n*-fold cross validation	111
Oligomer modeling	寡聚蛋白建模	181
PDB数据库	protein data bank, PDB	165

RNA测序	RNA-seq	468
RNA诱导的沉默复合物	RNA-induced silencing complex, RISC	409
S-腺苷甲硫氨酸	SAM	457
TN	true negative	112
TP	true positive	112
X-射线晶体分析法	X-ray diffraction crystallography	162
Z值	Z-score	194

A

癌基因组剖析计划	cancer genome anatomy project, CGAP	372
癌症差异甲基化区域	C-DMRs	473
安琪儿综合征	Angelman syndrome	466

B

贝-威二氏综合征	Beckwith-Wiedmann syndrome	466
比对界面	Alignment Interface	181
比较基因组学	comparative genomics	7, 233
比较模建法	comparative modeling	180
表达数量性状位点	expression quantitative trait loci, eQTL	392
表达序列标签	expressed sequence tag, EST	21, 43
表观遗传变异	epimutations	483
表观遗传学	epigenetics	456

C

参考序列	reference sequence	78
测试集	test set	111
层次聚类	hierarchical clustering	102
插入/删除多态	In/Del	376
差异甲基化的区域	differentially methylated region, DMR	466
差异甲基化区域	DMRs	473
差异甲基化杂交	DMH	459
缠绕法	wrapper method	92
超二级结构	supersecondary structure	155
超家族	super family	169
沉默子		289
重编程差异甲基化区域	R-DMRs	473
重抽样	re-sampling	111
重叠群	contig	43
重复序列	repetitive sequence	61
重亚硫酸盐测序	BS-seq	469

重亚硫酸盐测序技术	bisulphite conversion followed by capture and sequencing, BC-seq	461
重亚硫酸盐甲基化谱	bisulfite methylation profiling, BiMP	460
初始miRNA	primary RNA, pri-RNA	408
传递不平衡检验	transmission disequilibrium test, TDT	381
串联重复数据库	tandem repeats database, TRDB	28
磁共振	nuclear magnetic resonance, NMR	163
从头预测方法	Abinitio	180

D

代谢通路	metabolic pathway	337
代谢网络进化	metabolic network evolution	238
单核苷酸多态性	single nucleotide polymorphisms, SNPs	375
单核苷酸多态性	single-nucleotide polymorphism, SNP	8
单亲二倍体	UPD	467
单体	monomer	157
单元	singletion	228
蛋白互作对		236
蛋白质		505
蛋白质互作对	protein interaction pair	236
蛋白质结构分类数据库	structural classification of protein, SCOP	168
蛋白质结构域	protein domain	169
蛋白质组学	proteomics	11
等位	allele	375
等位基因特异性甲基化	ASM	489
点突变可接受矩阵	point accepted matrix, PAM	34
点阵	dot matrix	37
读段	read	424
短串联重复DNA数据库	short tandem repeat DNA internet database, STRBase	28
断裂基因	interrupted gene	56
对照细胞	control cell	76
多重命中	multiple hit	229
多基因病	ploygenic disorder	368
多梳家族蛋白	polycomb group protein, PcG	478
多维度标度技术	multi-dimensional scaling, MDS	187
多序列比对	multiple sequence alignment	43

E

二分网络	bipartite network	333
二级结构	secondary structure	154

F

发育差异甲基化区域	D-DMRs	473
发育重编程	developmental reprogramming	489
反馈环	feedback loop	444
反转录	reverse transcription	76
飞摩尔	femtomole	164
非编码RNA	non-coding RNA	424
非编码RNA数据库	noncoding RNA database	28
非翻译区	untranslated region, UTR	408
非共价键形式	non-covalent	198
非同义SNP	non-synonymous SNP	376
分层聚类法	hierarchical clustering	54
分类系统理论	system of taxonomy	145
分裂法	divisive	102
分子病	molecular disease	202
分子量	molecular weight, MW	164
分子钟	molecular clock	224
峰值探测	peak calling	464
负选择	negative selection	226
复合体	complex	203
复杂疾病	complex disease	368

G

高分值片段对	high-scoring pairs, HSPs	40
高通量组学	high-throughput omics	6
个体化治疗	personalized medicine	9
个体间的差异甲基化区域	Inter-DMRs	473
个体内的差异甲基化区域	Intra-DMRs	473
工程模式	Project mode	181
功能基因组学	functional genomics	5,8,118,233
供体位点	donor	58
共性序列	consensus sequence	43
共有序列	consensus sequence	288
寡核苷酸芯片	oligonucleotide microarray	77
关联研究	association study	380
光纤微珠芯片	beadarray microarray	79
国际人类单体型图计划	international HapMap project, HapMap	382
国际人类基因组单体型图计划	The International HapMap Project	5

过出现 over-presentation 140
过滤法 filter method 92

H

罕见疾病 rare disease 368
核酸 505
核糖体RNA rRNA 284
核心启动子 288
后基因组 post-genomics 118
后基因组学 post-genomics 5,8
互补对 complementary pairs 198
化学交联 cross-linking 164
环境基因组计划 environment genome project, EGP 369
荟萃标志物 Meta-signature 507

J

肌球蛋白 myoglobin 44
基因本体论 gene ontology 145
基因表达谱 gene expression profile 81
基因复制 gene duplication 32
基因富集分析 gene set enrichment analysis, GSEA 139
基因图谱 gene map 5
基因芯片 gene chip 75
基因型 genotype 375
基因组tRNA数据库 genomic tRNA database, GtRDB 28
基因组测序序列 genome survey sequences, GSS 21
基因组范围关联研究 genome-wide association study, GWAS 389
基因组功能注释 genome annotation 8, 118
基因组学 genomics 4
基于多维标度技术 MDS 194
基于相似性方法 similarity-based approaches 193
集富集分析 gene set enrichment analysis 479
加权网络 weighted network 333
家族 family 157
甲基化CpG岛扩增法 methylated CpG island amplification, MCA 459
甲基化DNA免疫共沉淀测定技术 MeDIP 458
甲基化测序 Methyl-seq 460
甲基化间区位点扩增 AIMS 459
甲基化敏感切割位点计数 methylation-sensitive cut counting, MsCC 460

甲基化敏感随机性引物PCR	methylation-sensitive arbitrarily primed PCR, MS-AP-PCR	459
甲基化数量性状位点	methQTLs	489
假基因数据库	PseudoGene	27
简捷模式	First Approach mode	181
简约信息位点	parsimony-informative site	223
简约重亚硫酸盐测序技术	reduced representation bisulphite sequencing, RRBS	461
酵母双杂交	yeast two Hybrid, Y2H	334
结点系谱	Term Lineage	122
结构风险最小化原则	structural risk minimization inductive principle, SRM	110
结构基因组学	structural genomics	233
结构域	domain	156, 157
介数	betweenness	343
进化创新	evolutionary innovation	226
进化基因组学	evolutionary genomics	233
进化论	theory of evolution	145
净化选择	purify selection	226
局部表面特征	local surface patterns, clefts	196
聚类分析	cluster analysis	98
聚类系数	clustering coeffcient	343
卷曲	coil	155
卷曲螺旋	coiled-coil	67
决策树	decision tree	108
均衡正确率	balanced accuracy	112

K

开放读码框	Open Reading Frame, ORF	56
拷贝数变异	copy number variants, CNV	376
拷贝数变异	copy number variant, CNV	8
可变模板结构	alternative template structures	173
空位	gap	32
空位罚分	gap penalty	33
空位扩展	gap extension	33
空位设置	gap opening	33

L

冷冻电子显微镜	cryoelectron microscopy	163

连锁块	linkage block	376
连通度或度	degree	342
联亲和纯化-质谱	tandem affinity purification – mass spectrometry, TAP-MS	334
镰状细胞贫血症	sickle-cell anemia	202
亮氨酸拉链	leucine zipper	156
留一法交叉证实	leave-one-out cross validation, LOOCV	112

M

美国国家生物技术信息中心	NCBI	20
美国国立生物技术信息中心	national center for biotechnology information, NCBI	113
免疫共沉淀	co-immunoprecipitation	334
敏感性	sensitivity	112
模糊功能结构	fuzzy functional forms, FFFS	193
模块化测度	modularity measure	349
模拟退火算法	simulated annealing algorithm	90
模式识别	pattern recognition	164
模体	motif	428
模体	Motif	58

N

凝聚法	agglomerative	102

O

欧洲分子生物学研究中心	EMBL	20

P

旁系同源	paralogy	32
旁系同源体	paralogs	32
配对比对	pairwise alignment	45
偏序图	partial-order graph	47
平均距离	average distance	344
评价准则	evaluation criteria	89
普瑞德威利综合征	Prader-Willi syndrome	466

Q

启动子		287

启发式搜索	heuristic search	90
前导肽	leader peptide	67
前馈环	Feed-forward Loop	444
前体miRNA	pre-miRNA	408
切割点	cut point, CP	440
全基因组鸟枪重亚硫酸盐测序	whole-genome shotgun bisulfate sequencing, WGSBS	461
全面高通量芯片相对甲基化技术	comprehensive high-throughput arrays for relative methylation, CHARM	460

R

染色质免疫共沉淀	ChIP	460
染色质免疫共沉淀-测序	chromatin immunoprecipitation sequencing, ChIP-seq	464
染色质免疫共沉淀-芯片	ChIP-chip	463
热点	hot spots	198
热图	heat map	104
人类表观基因组计划	HEP	458
人类基因组计划	human genome project, HGP	4,367
日本的DNA数据库	DDBJ	20

S

三维基序	3-dimensional motif-based	194
上位效应	epistasis	369
生物标记	biomarker	6
生物信息学	bioinformatics	2
生物学标记	biomark	418
实验细胞	experimental cell	76
受体位点	acceptor	58
树长	tree length	222
数据标准化	normalization	83
数量性状位点	quantitative trait loci, QTL	391
双向搜索	bi-direction search	89
双重的负反馈环	double-negative feedback	445
顺序	sequence	154
搜索策略	search strategy	89
随机缠绕	random coil	202
随机搜索	non-deterministic search	89, 90

T

贪婪登山法	greedy climbing hill	90
特异性	specificity	112
特征选择	feature selection	89
调控的反馈环	regulated feedback loop	445
同源建模	homology modeling	164
同源建模	homology-modelling	173
同源性	homology	32

W

完全搜索	complete search	90
网络	network	332
网络模块	network module	236,347
网络模体		236
网络模体	network motif	443
网络模序	motif	346
微处理器	microprocessor	408
微卫星	microsatellite, MS	376
微效性	minor effect	369
微阵列	microarray	75
位点长度	site length	222
位点构型	site configuration	223
位点模式	site pattern	223
位点频谱	site-frequency spectrum	228
无尺度	scale-free	237,436
无尺度网络	scale-free network	345
无规卷曲	random coil	155
无权网络	unweighted network	333
无向网络	undirected network	332
物理图谱	physical map	4

X

限制性内切酶或者限制性的标记的基因组扫描	RLGS	458
相对速率检验	relative-rate test	224
相互作用界面	interface	198
相互作用位点	interaction sites	198
相似性	similarity	32

镶嵌法	embedded method	92
向后选择	backward elimination	89
向前选择	forward selection	89
小干扰RNA	small interfering RNA, siRNA	408
小世界	small world	237
信号传导通路	signal transduction pathway	337
信号肽	signal peptide	67
信使RNA	mRNA	284
性状长度	character length	222
序列比对	sequence alignment	32
序列片段对	segment pair	40
序列图谱	sequence map	5
选择压力	selective pressure	43
血源一致性	identical-by-descent, IBD	379
训练集	training set	111

Y

亚基	subunit	156,203
阳性预测率	positive predictive value, precision	112
药物基因组学	pharmacogenomics	9
一级结构	primary structure	154
遗传关联数据库	genetic association database, GAD	371
遗传算法	genetic algorithm	90
遗传图谱	genetic map	4
遗传异质性	heterogeneity	369
异变的甲基化区域	VMRs	489
阴性预测率	negative predictive value	112
印记控制区	imprinting control region, ICR	466
有向网络	directed network	332
有向无环图	directed acyclic graph	47
预处理	pre-procession	81
原位合成芯片	light-controlled in situ synthesis of DNA microarrays	77

Z

增强子		288
折叠子	fold	169
真阳性		112
正确率	accuracy	112

正选择	positive selection	226
支持向量	support vector	110
支持向量机	support vector machine, SVM	110, 205
直径	diameter	344
直系同源	orthology	32
直系同源体	orthologs	32
指导树	guide tree	45
中间节点	intermediate nodes, ITN	440
中心节点	hub	342
中性学说	neutral theory of molecular evolution	224
种系形成	speciation	32
转换-颠换矩阵	transition-transversion matrix	34
转录因子		300
转录因子	transcriptional factor, TF	428
转录组学	transcriptomics	9
转运RNA	tRNA	284
状态一致性	identical-by-state, IBS	380
自组织映射	self-organizing map, SOM	104
组蛋白甲基转移酶	HMTs	478
组蛋白去甲基化酶	HDMTs	478
组蛋白去乙酰化酶	HDAC	478
组蛋白修饰	histone modification	461
组蛋白乙酰化转移酶	HAT	478
组件	module	157
组织差异甲基化区域	T-DMRs	469, 473
祖先重建	ancestral reconstruction	222
最大完全子图	clique	436
最简约重建	most parsimonious reconstruction	223
最小等位频率	minor allele frequency, MAF	375
最小二乘法	least-squares, LS	221
最小公共超图	minimal common supergraph	47
最小进化树	minimum evolution tree	222
最优超分面	optimal hyperplane	110